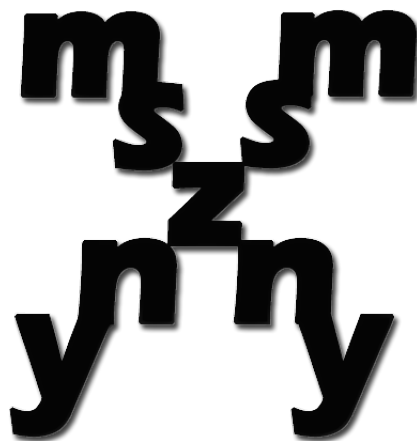


XVII. Magyar Számítógépes Nyelvészeti Konferencia



Szerkesztette:
Berend Gábor
Gosztolya Gábor
Vincze Veronika

Szeged, 2021. január 28–29.

Szerkesztette¹:

Berend Gábor, Gosztolya Gábor, Vincze Veronika
{berendg,ggabor,vinczev}@inf.u-szeged.hu

Felelős kiadó:

Szegedi Tudományegyetem
TTIK, Informatikai Intézet
6720 Szeged, Árpád tér 2.

ISBN: 978-963-306-781-9

Nyomtatta:

JATEPress
6722 Szeged, Petőfi Sándor sugárút 30–34.

Szeged, 2021. január

Az MSZNY 2021 konferencia szervezője:

MTA-SZTE Mesterséges Intelligencia Kutatócsoport

¹a L^AT_EX's 'confproc' csomagjára támaszkodva

Előszó

2021. január 28–29-én már tizenhetedik alkalommal kerül sor a Magyar Számítógépes Nyelvészeti Konferencia megrendezésére. Idén azonban rendhagyó módon, a virtuális térben tartjuk meg konferenciánkat, az ismert COVID-19 járványügyi helyzetre való tekintettel. Ugyanakkor bízunk benne, hogy a személyes találkozások és eszmecserék hiánya ellenére is sikeres és szakmailag mindenkit gazdagító eseménynek nézünk elébe.

A konferencia fő célkitűzése a kezdetek óta állandó: lehetőséget biztosítani a nyelv- és beszédtechnológia területén végzett kutatások eredményeinek ismertetésére és megvitatására, ezen felül a különféle hallgatói projektek, illetve ipari alkalmazások bemutatására. A hagyományokat követve a konferencia idén is nagyfokú érdeklődést váltott ki az ország nyelv- és beszédtechnológiai szakembereinek körében. A 32 beküldött cikkből gondos mérlegelést követően 26-ot fogadott el a programbizottság, melyek témája számos szakterületre terjed ki a beszédtechnológiai fejlesztésektől kezdve a legújabb nyelvi modellek bemutatásán keresztül a spontán beszéd elemzésére vonatkozó eredményekig.

Nagy örömet jelent számunkra, hogy Biszak Sándor és Biszak Előd elfogadták meghívásunkat, akik a digitális archívumok létrehozásával kapcsolatos tapasztalataikról fognak beszámolni plenáris előadásuk során.

Az idei évben is különdíjjal jutalmazzuk a konferencia legjobb cikkét, mely a legjelentősebb eredményekkel járul hozzá a magyarországi nyelv- és beszédtechnológiai kutatásokhoz. Ezen felül immár harmadik alkalommal osztjuk ki a legjobb bíráló díját, amellyel a bírálók fáradságos, ugyanakkor nélkülözhetetlen munkáját kívánjuk elismerni.

Köszönettel tartozunk az MTA-SZTE Mesterséges Intelligencia Kutatócsoportjának és a Szegedi Tudományegyetem Informatikai Intézetének helyi szervezésben segédkező munkatársainak. Végezetül szeretnénk megköszönni a programbizottság és a szervezőbizottság minden tagjának áldozatos munkáját, ami nélkül nem jöhetett volna létre a konferencia.

A szervezőbizottság nevében,

Ács Judit, Berend Gábor, Gosztolya Gábor, Novák Attila, Sass Bálint, Simon Eszter, Sztahó Dávid, Vincze Veronika

Tartalomjegyzék

Nyelvmodellek	1
3	Introducing huBERT <i>Dávid Márk Nemeskey</i>
15	Evaluating Contextualized Language Models for Hungarian <i>Judit Ács, Dániel Lévai, Dávid Márk Nemeskey, András Kornai</i>
29	HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben <i>Feldmann Ádám, Váradi Tamás, Hajdu Róbert, Indig Balázs, Sass Bálint, Makrai Márton, Mittelholcz Iván, Halász Dávid, Yang Zijian Győző</i>
Transzkripció, transliteráció	37
39	Magyar hadifoglyok adatainak orosz-magyar átírása és helyreállítása, és a szabadszöveges adatbázisok tulajdonságai <i>Sass Bálint, Mittelholcz Iván, Halász Dávid, Lipp Veronika, Kalivoda Ágnes</i>
53	emPhon: Morphologically sensitive open-source phonetic transcriber <i>Kulcsár Virág, Lévai Dániel</i>
63	Automatic punctuation restoration with BERT models <i>Nagy Attila, Bial Bence, Ács Judit</i>
Szemantika	75
77	Mitigating the Knowledge Acquisition Bottleneck for Hungarian Word Sense Disambiguation using Multilingual Transformers <i>Gábor Berend</i>
91	Analysing the semantic content of static Hungarian embedding spaces <i>Tamás Ficsor, Gábor Berend</i>
107	Interaktív tematikus-szemantikus térkép a Történeti Magánéleti Korpusz keresőfelületén <i>Novák Attila</i>
Beszédtechnológia	121
123	3D konvolúciós neuronhálón és neurális vokóderen alapuló néma-beszéd-interfész <i>Tóth László, Amin Shandiz, Gosztolya Gábor, Zainkó Csaba, Markó Alexandra, Csapó Tamás Gábor</i>

- 139 End-to-end és hibrid mélyneuronháló alapú gépi leiratozás magyar nyelvű telefonos ügyfélszolgálati beszélgetésekre
Mihajlik Péter, Balog András, Tarján Balázs, Fegyő Tibor
- 147 Enyhe kognitív zavar detektálása beszédhangból x-vektor reprezentáció használatával
José Vicente Egas-López, Balogh Réka, Imre Nóra, Tóth László, Vincze Veronika, Pákáski Magdolna, Kálmán János, Hoffmann Ildikó, Gosztolya Gábor
- 157 FORvoice 120+: Statisztikai vizsgálatok és automatikus beszélő verifikációs kísérletek időben eltérő felvételek és különböző beszéd feladatok szerint
Sztahó Dávid, Beke András, Szaszák György

Spontán beszéd, chat

167

- 169 A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata
Vincze Veronika, Üveges István, Szabó Martina Katalin, Takács Károly
- 183 Magyar nyelvű spontán beszéd szemantikai-pragmatikai sajátosságainak elemzése nagy méretű korpusz (StaffTalk) alapján
Vincze Veronika, Üveges István, Szabó Martina Katalin
- 197 Egy nyílt forráskódú magyar időpont-egyeztető chatbot
Nagy Soma Bálint, Herdina Viktor, Farkas Richárd

Poszter, laptopos bemutató

209

- 211 StaffTalk: magyar nyelvű spontán beszélgetések korpusza
Szabó Martina Katalin, Vincze Veronika, Ring Orsolya, Üveges István, Vit Eszter, Samu Flóra, Gulyás Attila, Galántai Júlia, Svetelszky Zsuzsanna, Bodor-Eranus Eliza Hajnalka, Takács Károly
- 225 Automatikus írásjelek visszaállítása és Nagybetűsítés statikus korpuszon, transzformer modellen alapuló neurális gépi fordítással
Yang Zijian Győző
- 233 Smooth inverse frequency based text data selection for medical dictation
Domonkos Bálint, Péter Mihajlik
- 243 Automatikus hibajavítás statikus szövegeken
Máté Gulás, Yang Zijian Győző, Andrea Dömötör, László János Laki
- 253 Szó, beszéd – avagy hogyan kommunikálunk egymásról
Üveges István, Szabó Martina Katalin, Vincze Veronika

- 265 Egy következtetésvezérelt csevegőrobot anatómiája. Az ITSy-Bitsy modell
Kilián Imre
- 275 A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei
Vincze Veronika, Kicsi András, Főző Eszter, Vidács László

Szintaxis, szemantika

289

- 291 Jogi szövegek tezaurusz alapú osztályozása: egy nyelvfüggetlen modell létrehozásának problémái
Nyéki Bence
- 305 Egy nagyobb magyar UD korpusz felé
Novák Attila, Novák Borbála
- 319 Értsük meg a magyar entitás-felismerő rendszerek viselkedését!
Farkas Richárd, Nemeskey Dávid Márk, Zahorszki Róbert, Vincze Veronika

Szerzői index, névmutató

331

NYELVMODELLEK

Introducing huBERT

Nemeskey Dávid Márk¹

¹Számítástechnikai és Automatizálási Kutatóintézet
nemeskey.david@gmail.com

Abstract. This paper introduces the huBERT family of models. The flagship is the eponymous BERT Base model trained on the new Hungarian Webcorpus 2.0, a 9-billion-token corpus of Web text collected from the Common Crawl. This model outperforms the multilingual BERT in masked language modeling by a huge margin, and achieves state-of-the-art performance in named entity recognition and NP chunking. The models are freely downloadable.

Keywords: huBERT, BERT, evaluation, NER, chunking, masked language modeling

1 Introduction

Contextualized embeddings, since their introduction in McCann et al. (2017) have altered the natural language processing (NLP) landscape completely. Systems based on ELMo (Peters et al., 2018), and especially BERT (Devlin et al., 2019) have improved the state of the art for a wide range of benchmark tasks. The improvement is especially notable for high-level natural language understanding (NLU) tasks, such as the ones that make up the GLUE (Wang et al., 2018) and SQuAD (Rajpurkar et al., 2016, 2018) datasets. In the long run, BERT proved more successful than ELMo, not least because once it has been *pretrained* on large amount of texts, it can be *finetuned* on any downstream task, while ELMo cannot stand on its own and must be integrated into traditional NLP systems.

The triumph of BERT also marks the move away from LSTMs (Hochreiter and Schmidhuber, 1997) toward the attention-based Transformer (Vaswani et al., 2017) architecture as the backbone of language representation models. BERT was soon followed by an abundance of similar models, such as RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) or BART (Lewis et al., 2019). These models tweak different aspects of BERT, including the amount of training data, the tasks used to pretrain it, or the architecture itself. Each paper reports improvements over the last.

As always in NLP¹, all the pioneering research above was centered on English. Support for other languages came in two forms: native contextual embeddings, such as CamemBERT (Martin et al., 2019) for French, or multilingual variations of the models above. Examples for the latter are multi-BERT and XLM-RoBERTa (Conneau et al., 2019), both of which were trained on corpora with around 100 languages (Wikipedia for the former, the Common Crawl² for latter).

¹ With the possible exception of morphology.

² <https://commoncrawl.org/>

Both alternatives have their own advantages and disadvantages: native models are as expensive to train as the original English ones, costing up to hundreds of thousands of euros; which seems excessive, especially since good quality multilingual models have already been published. On the other hand, the capacity of multilingual models is shared among the many languages they support, which hurts single-language performance. Medium-size languages, of which Hungarian is one, are further disadvantaged by the size of the available textual data. In the training corpora of multilingual models, larger languages are represented by a proportionally higher amount of text, which introduces serious bias into the final models. Taking this all into consideration, we came to the conclusion that Hungarian is probably better served by native models.

In this paper, we introduce the `huBERT` family of models. As of now, the family consists of two preliminary BERT Base models trained on Wikipedia and the eponymous `huBERT` model, trained on a new nine-billion-token corpus; it is also the first publicly available Hungarian BERT model. We evaluate `huBERT` against multi-BERT on the two tasks they were pretrained, as well as on two downstream tasks: named entity recognition (NER) and NP chunking. We find that `huBERT` outperforms multi-BERT on the training tasks by a huge margin, and achieves a new state of the art in both NER and NP chunking, thereby strengthening our concluding sentence in the last paragraph.

The rest of the paper is organized as follows. In Section 2, we describe the training corpora and the pretraining process behind the models. Section 3 details our experimental setup and presents our results. Section 4 highlights a few shortcomings of relying solely on the new contextualized embedding machinery. Finally, we conclude our work in Section 5.

2 Pretraining

In this section we describe the pretraining procedure in detail in the hope that it helps others embarking on a similar venture avoiding potential pitfalls along the way.

2.1 Background

Pretraining modern contextualized representations is a costly business. The models themselves are huge (BERT-Base, which has become a standard, has 110 million parameters), and the associated training corpora also start at several billion words. The quadratic resource requirements of the attention mechanism can only be accommodated by high-end hardware. These factors all add up, and as a result, training a modern Transformer model takes days or weeks on hundreds of GPUs or TPUs.

The financial costs incurred by such a training regimen are prohibitive for smaller laboratories, unless they receive support from the industry. However, most of the time the support is limited and it does not allow experimentation with model architectures, let alone hyperparameter tuning. This means that pretraining for smaller groups is a leap of faith, which either succeeds or not; and this inequality of the playing field raises various ethical issues (Parra Escartín et al., 2017).

Our situation is not different. We were kindly given the use of 5 v3-8 TPUs by Google in the Tensorflow Research Cloud (TFRC)³ program, as well as two weeks on a v3-256 TPU Pod. Our main goal was to train a BERT-Base model on Webcorpus 2.0: a new, 9-billion-token corpus compiled from the Hungarian subset of the Common Crawl (Nemeskey, 2020b). Based on the numbers in the original BERT paper, we calculated that two weeks should be enough to train the model to convergence. However, an earlier failed attempt at pretraining an ALBERT (Lan et al., 2019) model that never converged convinced us to start with a smaller corpus to ensure that the training process works.

2.2 huBERT Wiki

At about 170 million words in 400 thousand documents⁴, the Hungarian Wikipedia is but a fraction of the English one. After filtering it according to the BERT guidelines, its size further decreases to about 110 million words in 260 thousand documents. This is considerably smaller than Webcorpus 2.0, but it contains good quality, edited text, which makes it a valuable training resource. Its small size also allowed us to pretrain a BERT-Base model on it in 2.5 days on a single v3-8 TPU.

BERT models usually come in two flavors: *cased* and *uncased*. The former operates on unprocessed text; in the latter, tokens are lower cased and diacritical marks are removed. In keeping with this practice, we also trained two variants. However, as diacritics are *distinctive* in Hungarian, we could not afford to lose them, and replaced the uncased model with a *lower cased* one.

BERT models are pretrained with two tasks: *masked language modeling (MLM)* and *next sentence prediction (NSP)*. The language understanding capabilities of the model reportedly derive from the former (Lan et al., 2019; Liu et al., 2019), as NSP is very easy to learn. Since we used the original BERT training code, we kept both tasks.

As is the case with the English BERT, our models are all pretrained on sequences of up to 512 wordpieces. As the training configurations published in the literature are for much larger corpora, they are not directly adaptable to our case. Hence, we experimented with different training regimens for both the cased and lower cased variants:

1. Three models were trained with full-length sequences for 50,000, 100,000 and 200,000 steps. These roughly correspond to 90, 180 and 360 epochs, respectively;
2. Following the recommendation in the BERT GitHub repository, one model was trained with a sequence length of 128 for 500,000 steps (600 epochs) and with a sequence length of 512 for an additional 100,000 steps (or 180 epochs).

All models were trained with a maximum learning rate of 10^{-4} and the maximum possible batch size: 1024 for the model with 128-long sequences and 384 for the rest. The training data for the masked language modeling task was duplicated 40 times with different mask positions. The official training code uses a learning rate decay with a warmup period, which we set to 10% of the total number of training steps. The code unfortunately does not support early stopping; it does not even accept a validation set.

³ <https://www.tensorflow.org/tfrc>

⁴ 2018 snapshot

However, as we shall see in Section 3.1, performance on the test set showed no sign of overfitting.

All models use a wordpiece vocabulary of around 30,000 tokens to match the English BERT-Base models. Increasing it 5,000 tokens did not yield any improvements, so we opted for the smaller vocabulary in order to keep the model smaller.

Model	Seq. length	Steps	Hours	Masked LM	Next sentence
Cased	512	50,000	13	0.5544961	0.97125
	128	500,000	59	0.6669028	0.995
	512	+100,000	25	0.6657926	0.99
Lower	512	50,000	13	0.5538445	0.9825
	512	100,000	25	0.6100383	0.9975
	512	200,000	50	0.6273391	0.9975
	128	500,000	59	0.6425686	0.99125
	512	+100,000	25	0.665879	0.9975

Table 1. Training times and accuracies of the different BERT models on the two training tasks

Table 1 compares all configurations. In the cased case, the TPU went down for maintenance during training, so the 100,000 and 200,000-step models are missing from the results. Even without them, several observations can be made. First, the 50,000-step models clearly underfit the data, even though they were trained for twice as many epochs as the English BERT. On the other hand, the difference between the 100,000 and 200,000-step models is much smaller than between the 50,000 and 100,000-step models, suggesting a performance peak around 300,000–400,000 steps.

Second, in line with the findings of Lan et al. (2019); Liu et al. (2019), the next sentence prediction task seems very easy, as all but the first models attain over 99% accuracy. In contrast, the masked LM task proved much harder, and its accuracy seems rather low. Unfortunately, the evaluation results for the English BERT are not published anywhere, which makes it difficult to put the numbers in context. Based on the diminishing returns, the longest-trained models are likely to be close to the maximum achievable on Wikipedia alone.

Finally, our experiences confirmed that the two-stage training regimen recommended in the BERT repository indeed leads to better results. The rationale behind this method is that the first phase trains most of the model weights and the second phase is “*mostly needed to learn positional embeddings, which can be learned fairly quickly*”⁵. While this seems to be the case for the cased model, the masked LM accuracy of the lower cased model improved by more than 2% in the second phase, indicating either that sub-

⁵ <https://github.com/google-research/BERT/#pre-training-tips-and-caveats>

stantial learning still happens at this stage or that some of the dependencies in the data can be better exploited by a 512-token window.

2.3 huBERT

Having confirmed that the BERT training code works and produces functional models on Wikipedia, we proceeded to train the main huBERT model on the much larger Webcorpus 2.0. We used the same configuration as for the preliminary models, with two notable exceptions.

First, we only had time to pretrain one model. We chose to focus on the cased model, as that is more universally usable. Second, as opposed to single TPUs, TPU Pods are always preemptible, and our earlier experience with ALBERT taught us that the training might be interrupted several times a day. Unfortunately, the original BERT training script is not prepared for this eventuality and once interrupted, it can never resume training. To mitigate this issue, we wrote a wrapper script around the BERT training code that monitors the log file and restarts training whenever the TPU Pod goes down. We also decreased the number of steps between checkpoints to 1,000 (from the default 5,000) to minimize the work lost.

In the end, our training quota expired after 189,000 steps, cutting the pretraining slightly short. To validate the model, we ran the same evaluations as we did for the preliminary models, this time on a held-out portion of Webcorpus 2.0. The results (MLM accuracy of 0.63 with a sequence length of 128 and 0.66 with 512) closely follows those reported in Table 1, which indicates that the model might similarly be close to convergence and better results could only be expected of larger (e.g. BERT-Large) models.

2.4 Availability

All huBERT models can be downloaded freely from the huBERT homepage⁶. The main huBERT model is also available from the Hugging Face model repository⁷ under the moniker SZTAKI-HLT/hubert-base-cc.

The emBERT NER and NP taggers, described in Section 3.2, replace the original models based on multi-BERT and can be downloaded from inside emtsv or from the GitHub repository⁸.

3 Evaluation

BERT models are usually evaluated on high-level natural language understanding tasks, such as question answering or textual entailment. Unfortunately, no Hungarian benchmark datasets exist for these tasks. Because of this, we evaluate our models by contrasting their performance to the multi-language version of BERT in two ways:

⁶ <https://hlt.bme.hu/en/resources/hubert>

⁷ <https://huggingface.co/SZTAKI-HLT/hubert-base-cc>

⁸ <https://github.com/dlt-rilmta/emBERT-models>

1. We compare their accuracy on the two training tasks on a held-out portion of Wikipedia and Webcorpus 2.0.
2. We include our models in the `emBERT` module (Nemeskey, 2020a) and measure their performance on named entity recognition and NP chunking.

3.1 Training tasks

Table 2 presents the results of the first experiment. Both our cased and lower cased models achieve similar accuracies on the held-out set as on the training data, allaying any suspicion of overfitting. The `huBERT` Wiki models perform slightly better on Wikipedia than `huBERT`, but attain significantly lower accuracy on Webcorpus 2.0. Compared to this, `huBERT` is fairly robust across both corpora, no doubt benefiting from its much larger and more varied training corpus. All cased models clearly outperform multi-BERT on both tasks (multi-BERT is only available in the cased configuration).

Case	Model	Wikipedia		Webcorpus 2.0	
		MLM	NSP	MLM	NSP
Cased	multi-BERT	<i>0.00001</i>	<i>0.560</i>	<i>0.000004</i>	<i>0.455</i>
	<code>huBERT</code> Wiki	0.65	0.988	0.46	0.786
	<code>huBERT</code>	0.64	0.985	0.61	0.959
Lower	<code>huBERT</code> Wiki	0.641	0.99		

Table 2. Accuracy of multi-language BERT and members of the `huBERT` family on the two training tasks on the held-out set of the two training corpora.

In fact, the performance of multi-BERT leaves a lot to be desired. Its accuracy on the next sentence prediction task is, at 50%, effectively random. The masked LM loss is equivalent to a perplexity of about 130,000, which, given its vocabulary of 120,000 wordpieces, is even worse than that.

On the one hand, this abysmal performance comes as a surprise, for two reasons: first, it was also trained on Wikipedia; and second, multi-BERT fares much better on downstream tasks (see Section 3.2, below). On the other, it goes to show that multi-language models sacrifice too much of single-language performance to be of actual use for the tasks they were trained on. This underlines the importance of native Hungarian contextual embeddings.

3.2 NLP tasks

Tables 3 and 4 show the performance of `huBERT`-based models against leading Hungarian systems on NP chunking and NER, respectively⁹. The tables are extended versions

⁹ For training details and a more thorough description of the tasks and the corresponding data, the reader is referred to Nemeskey (2020a)

of those found in Nemeskey (2020a). One difference to note is that, for the sake of a fair comparison, we only included systems in Table 4 that were trained and tested on the standard split of the Szeged NER corpus.

Table 3 demonstrates that BERT-based models in general perform favorably compared to traditional statistical models, represented here by members of the `hunchunk` family. `multi-BERT` already outperforms `HunTag3` in maximal NP-chunking by 1.5% F1 score on the test set, but it could only match `hunchunk`’s results on minimal NPs. `huBERT Wiki`, on the other hand, improves both scores by 1–1.5%. `huBERT` tops the list with another 0.5% increase on both tasks, achieving a new state of the art on both.

System	Minimal	Maximal
<code>hunchunk/HunTag</code> (Recski, 2010)	95.48%	89.11%
<code>HunTag3</code> (Endrédy and Indig, 2015)	–	93.59%
<code>emBERT w/ multi-BERT</code>	95.58%	95.05%
<code>emBERT w/ huBERT Wiki</code>	96.64%	96.41%
<code>emBERT w/ huBERT</code>	97.14%	96.97%

Table 3. Comparison of Hungarian NP chunkers

The results for named entity recognition (see Table 4) are less straightforward. `emBERT` with `multi-BERT` achieves 1% higher F1 score than the previous best (Simon, 2013). As opposed to the NP chunking tasks, `huBERT Wiki` could not improve on the multilingual model – in fact, it reaches a slightly lower F1 score, even though the difference is not significant. `huBERT`, however, again manages to squeeze another 0.5% out of the data, setting a new record on the Szeged NER corpus.

System	F1
(Szarvas et al., 2006)	94.77%
<code>hunner</code> (Varga and Simon, 2007)	95.06%
<code>hunner</code> (Simon, 2013)	96.10%
<code>emBERT w/ multi-BERT</code>	97.08%
<code>emBERT w/ huBERT Wiki</code>	97.03%
<code>emBERT w/ huBERT</code>	97.62%

Table 4. Comparison of Hungarian NER taggers

4 All that glitters is not gold

In this section, we dive briefly behind the numbers and show that even though our BERT models established new state of the art on two downstream benchmarks, their actual behavior on real-world data might lack in some areas. It must be pointed out that the two issues described below occur only to the named entity tagger, which implies a problem with insufficient training data (see Nemeskey (2020a)) rather than with the capabilities of the model architecture itself.

4.1 Invalid tag sequences

The numbers for both NP chunking and NER paint a similar picture: all BERT-based taggers outperform traditional machine learning systems on both tasks, with `huBERT` beating `multi-BERT` by a few percent. In case of NER, the gap is as small as 0.5%, which hardly justifies spending the resources needed to train a native Hungarian BERT model. However, when the taggers are applied to data outside the Szeged NER corpus, a different picture emerges.

In the original `emBERT` system, the labels emitted by the taggers were output as-is. This runs the risk of producing invalid tag sequences, of which an example is shown in Table 5. Here, `multi-BERT` generates invalid sequences such as `B-ORG B-ORG`, `E-MISC E-MISC` and even `B-MISC I-PER`. The tag sequence emitted by `huBERT` Wiki also contains an error, and its classification is not better than `multi-BERT`’s, either. `huBERT`’s output, on the other hand, is perfectly valid and the tagging is much more accurate as well.

It is worth mentioning that invalid tag sequences are rare, as the attention mechanism BERT is based on is able to use information from all tokens in the sequence, and hence the model finds the boundaries of named entities most of the time. It is only when the input sentence has an odd structure that we encountered invalid tag sequences. Indeed, the sentence in Table 5 is not a sentence in the grammatical sense; instead, it is the list of characters in a play, mistakenly grouped together by the sentence splitter. Still, tokenization errors and fragmented data crop up in all corpora, and our systems have to be robust enough to handle them. The `huBERT`-based tagger can be more robust to unfamiliar input than the other two because it was trained solely on (large and often fragmented) Hungarian data.

Nevertheless, we cannot be sure that `huBERT` taggers always generate valid output and hence we implemented a Viterbi-like algorithm on top of the tagger that prevents invalid tag transitions. The transition probabilities are uniform for each valid transition between tags (i.e. `B-PER` \rightarrow `I-PER`) and 0 otherwise. We decided against learning the probabilities from the training corpus, as it would downweight rarely seen but otherwise valid transitions. This would effectively prevent us from correctly tagging `1-*` entities, as the `O` \rightarrow `O` transition is much more probable than `O` \rightarrow `1-MISC`, etc.

Sentence	multi-BERT	huBERT Wiki	huBERT	m-B Viterbi
BARABÁS	B-PER	B-ORG	B-PER	B-PER
ÁDÁMNÉ	E-PER	E-ORG	E-PER	E-PER
az	O	O	O	O
édesanyja	O	O	O	O
A	B-ORG	O	O	O
MESTER	B-ORG	B-ORG	B-ORG	B-ORG
SZTELLA	E-ORG	E-ORG	E-ORG	E-ORG
a	O	O	O	O
partnernője	O	O	O	O
MISI	B-MISC	1-MISC	1-PER	B-MISC
bohóc	E-MISC	O	O	E-MISC
NOVOTNI	B-MISC	B-PER	B-PER	B-MISC
NÁNÁSI	I-MISC	I-MISC	I-PER	I-MISC
PIRI	E-MISC	E-PER	E-PER	E-MISC
lektor	E-MISC	O	O	O
MAROSI	1-MISC	1-MISC	1-MISC	1-MISC
újságíró	O	O	O	O
LITTKÉNÉ	B-MISC	B-MISC	1-PER	B-PER
NÉGY	I-PER	I-MISC	O	I-PER
KATONA	I-PER	I-MISC	O	I-PER
PERECESLÁNY	E-PER	E-MISC	O	E-PER

Table 5. Invalid tag sequences on a text fragment from the screenplay of Tragédia (1979) by István Örkény

As seen in the last column of Table 5, applying the Viterbi algorithm to the class transitions prevents the emission of invalid tag sequences, and occasionally improves the results as well.

4.2 Overenthusiasm

Applying the NER taggers to single words demonstrates another peculiarity of our BERT-based taggers: they seem overly enthusiastic to give a non-O label to almost any single word, including “a” (the), “macska” (cat) or “fut” (run). This does not happen when the words are in a sentential context, e.g. “a macska fut” (the cat is running) is correctly tagged as O O O. The cause of this behavior is not yet clear, as the training corpus contains no one-word “sentences”, and thus requires further research. As mentioned above, the chunker models are unaffected by this issue, which makes the NER training corpus the primary suspect.

5 Conclusion and future work

In this paper, we have introduced the huBERT family of models. The first three members of the family are two preliminary BERT-Base models pretrained on Wikipedia and

the eponymous huBERT model pretrained in Webcorpus 2.0. According to our tests, all models, but especially the latter, outperform the multilingual BERT model both in the tasks used to pretrain them and in token classification tasks, such as NP chunking and NER. huBERT achieves a new state of the art in both NLP tasks. Additionally, the models trained on solely Hungarian corpora seemed more stable when applied to unfamiliar text. huBERT is available on the Hugging Face Model Hub in both Pytorch and TensorFlow flavors.

In the future, we expect further, more recent models, such as Electra (Clark et al., 2020), to be added to the family.

Acknowledgements

This work was partially supported by National Research, Development and Innovation Office (NKFIH) grants #115288: “*Algebra and algorithms*” and #120145: “*Deep Learning of Morphological Structure*”, as well as by National Excellence Programme 2018-1.2.1-NKP-00008: “*Exploring the Mathematical Foundations of Artificial Intelligence*”.

huBERT was trained on TPUs provided by the Tensorflow Research Cloud program. Their support is gratefully acknowledged.

The authors thank Eszter Simon for bringing the issue of the emBERT NER model’s outputting invalid tag sequences to their attention and for the anonymous reviewers for their valuable insights.

Bibliography

- Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: Electra: Pre-training text encoders as discriminators rather than generators. In: 8th International Conference on Learning Representations, ICLR 2020 (2020), <https://openreview.net/forum?id=r1xMH1BtvB>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2019)
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of NAACL (2019)
- Endrédi, I., Indig, B.: HunTag3, a General-purpose, Modular Sequential Tagger – Chunking Phrases in English and Maximal NPs and NER for Hungarian, p. 213–218. Uniwersytet im. Adama Mickiewicza w Poznaniu, Poznan (2015)
- Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (11 1997)
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. CoRR abs/1909.11942 (2019), <http://arxiv.org/abs/1909.11942>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach (2019)
- Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. arXiv preprint arXiv:1911.03894 (2019)
- McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in translation: Contextualized word vectors. In: Advances in Neural Information Processing Systems. pp. 6294–6305 (2017)
- Nemeskey, D.M.: Egy `emBERT` próbáló feladat. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). pp. 409–418. Szeged (2020a)
- Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020b)
- Parra Escartín, C., Reijers, W., Lynn, T., Moorkens, J., Way, A., Liu, C.H.: Ethical considerations in NLP shared tasks. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing. pp. 66–73. Association for Computational Linguistics, Valencia, Spain (04 2017), <https://www.aclweb.org/anthology/W17-1608>
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
- Rajpurkar, P., Jia, R., Liang, P.: Know what you don’t know: Unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 784–789. Association for Computational Linguistics, Melbourne, Australia (07 2018), <https://www.aclweb.org/anthology/P18-2124>
- Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (11 2016), <https://www.aclweb.org/anthology/D16-1264>
- Recski, G.: Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel. In: Tanács, A., Vincze, V. (eds.) VII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 333–341 (2010)
- Simon, E.: Approaches to Hungarian Named Entity Recognition (2013), ph.D. Thesis, Budapest University of Technology and Economics
- Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms. In: Discovery Science, 9th International Conference, DS 2006, Barcelona, Spain, October 8-10, 2006, Proceedings. pp. 268–278 (2006)
- Varga, D., Simon, E.: Hungarian named entity recognition with a maximum entropy approach. Acta Cybern. 18(2), 293–301 (Feb 2007)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran As-

sociates, Inc. (2017), <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2018)

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems. pp. 5754–5764 (2019), <https://papers.nips.cc/paper/8812-xlnet-generalized-autoregressive-pretraining-for-language-understanding.pdf>

Evaluating Contextualized Language Models for Hungarian

Judit Ács^{1,2}, Dániel Lévai³, Dávid Márk Nemeskey², András Kornai²

¹ Department of Automation and Applied Informatics
Budapest University of Technology and Economics

² Institute for Computer Science and Control

³ Alfréd Rényi Institute of Mathematics

Abstract. We present an extended comparison of contextualized language models for Hungarian. We compare huBERT, a Hungarian model against 4 multilingual models including the multilingual BERT model. We evaluate these models through three tasks, morphological probing, POS tagging and NER. We find that huBERT works better than the other models, often by a large margin, particularly near the global optimum (typically at the middle layers). We also find that huBERT tends to generate fewer subwords for one word and that using the last subword for token-level tasks is generally a better choice than using the first one.

Keywords: huBERT, BERT, evaluation

1 Introduction

Contextualized language models such BERT (Devlin et al., 2019) drastically improved the state of the art for a multitude of natural language processing applications. Devlin et al. (2019) originally released 4 English and 2 multilingual pretrained versions of BERT (mBERT for short) that support over 100 languages including Hungarian. BERT was quickly followed by other large pretrained Transformer (Vaswani et al., 2017) based models such as RoBERTa (Liu et al., 2019b) and multilingual models with Hungarian support such as XLM-RoBERTa (Conneau et al., 2019). Huggingface released the Transformers library (Wolf et al., 2020), a PyTorch implementation of Transformer-based language models along with a repository for pretrained models from community contribution ¹. This list now contains over 1000 entries, many of which are domain- or language-specific models.

Despite the wealth of multilingual and language-specific models, most evaluation methods are limited to English, especially for the early models. Devlin et al. (2019) showed that the original mBERT outperformed existing models on the XNLI dataset (Conneau et al., 2018b). mBERT was further evaluated by Wu and Dredze (2019) for 5 tasks in 39 languages, which they later expanded to over 50 languages for part-of-speech tagging, named entity recognition and dependency parsing (Wu and Dredze, 2020).

¹ <https://huggingface.co/models>

Nemeskey (2020) released the first BERT model for Hungarian named *huBERT* trained on Webcorpus 2.0 (Nemeskey, 2020, ch. 4). It uses the same architecture as BERT base with 12 Transformer layers with 12 heads and 768 hidden dimension each with a total of 110M parameters. huBERT has a Word-Piece vocabulary with 30k subwords.

In this paper we focus on evaluation for the Hungarian language. We compare huBERT against multilingual models using three tasks: morphological probing, POS tagging and NER. We show that huBERT outperforms all multilingual models, particularly in the lower layers, and often by a large margin. We also show that subword tokens generated by huBERT’s tokenizer are closer to Hungarian morphemes than the ones generated by the other models.

2 Approach

We evaluate the models through three tasks: morphological probing, POS tagging and NER. Hungarian has a rich inflectional morphology and largely free word order. Morphology plays a key role in parsing Hungarian sentences.

We picked two token-level tasks, POS tagging and NER for assessing the sentence level behavior of the models. POS tagging is a common subtask of downstream NLP applications such as dependency parsing, named entity recognition and building knowledge graphs. Named entity recognition is indispensable for various high level semantic applications.

2.1 Morphological probing

Probing is a popular evaluation method for black box models. Our approach is illustrated in Figure 1. The input of a probing classifier is a sentence and a target position (a token in the sentence). We feed the sentence to the contextualized model and extract the representation corresponding to the target token. We use either a single Transformer layer of the model or the weighted average of all layers with learned weights. We train a small classifier on top of this representation that predicts a morphological tag. We expose the classifier to a limited amount of training data (2000 training and 200 validation instances). If the classifier performs well on unseen data, we conclude that the representation includes said morphological information. We generate the data from the automatically tagged Webcorpus 2.0. The target words have no overlap between train, validation and test, and we limit class imbalance to 3-to-1 which resulted in filtering some rare values. The list of tasks we were able to generate is summarized in Table 1.

2.2 Sequence tagging tasks

Our setup for the two sequence tagging tasks is similar to that of the morphological probes except we train a shared classifier on top of all token representations. Since multiple subwords may correspond to a single token (see Section 3.1 for

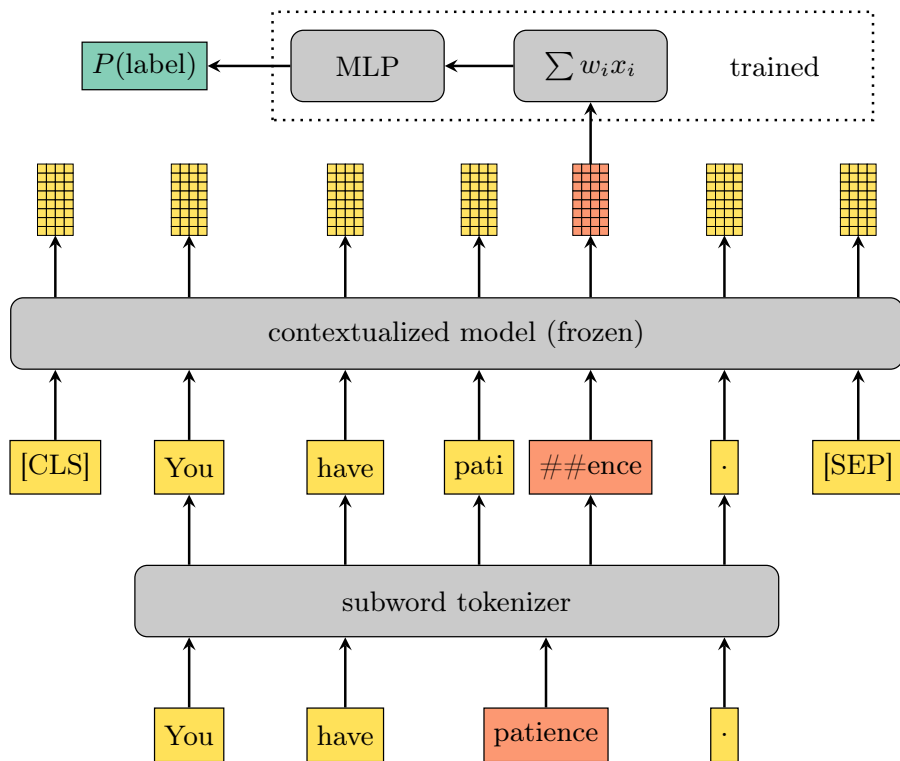


Fig. 1: Probing architecture. Input is tokenized into subwords and a weighted average of the mBERT layers taken on the last subword of the target word is used for classification by an MLP. Only the MLP parameters and the layer weights w_i are trained.

more details), we need to aggregate them in some manner: we pick either the first one or the last one.²

We use two datasets for POS tagging. One is the Szeged Universal Dependencies Treebank (Farkas et al., 2012; Nivre et al., 2018) consisting of 910 train, 441 validation, and 449 test sentences. Our second dataset is a subsample of Webcorpus 2 tagged with emtsv (Indig et al., 2019) with 10,000 train, 2000 validation, and 2000 test sentences.

Our architecture for NER is identical to the POS tagging setup. We train it on the Szeged NER corpus consisting of 8172 train, 503 validation, and 900 test sentences.

² We also experimented with other pooling methods such as elementwise max and sum but they did not make a significant difference.

Morph tag	POS	#classes	Values
Case	noun	18	Abl, Acc, . . . , Ter, Tra
Degree	adj	3	Cmp, Pos, Sup
Mood	verb	4	Cnd, Imp, Ind, Pot
Number psor	noun	2	Sing, Plur
Number	adj	2	Sing, Plur
Number	noun	2	Sing, Plur
Number	verb	2	Sing, Plur
Person psor	noun	3	1, 2, 3
Person	verb	3	1, 2, 3
Tense	verb	2	Pres, Past
VerbForm	verb	2	Inf, Fin

Table 1. List of morphological probing tasks.

2.3 Training details

We train all classifiers with identical hyperparameters. The classifiers have one hidden layer with 50 neurons and ReLU activation. The input and the output layers are determined by the choice of language model and the number of target labels. This results in 40k to 60k trained parameters, far fewer than the number of parameters in any of the language models.

All models are trained using the Adam optimizer (Kingma and Ba, 2014) with $lr = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We use 0.2 dropout for regularization and early stopping based on the development set.

3 The models evaluated

We evaluate 5 models.

huBERT the Hungarian BERT, is a BERT-base model with 12 Transformer layers, 12 attention heads, each with 768 hidden dimensions and a total of 110 million parameters. It was trained on Webcorpus 2.0 (Nemeskey, 2020), 9-billion-token corpus compiled from the Hungarian subset of Common Crawl³. Its string identifier in Huggingface Transformers is `SZTAKI-HLT/hubert-base-cc`.

mBERT the cased version of the multilingual BERT. It is a BERT-base model with identical architecture to huBERT. It was trained on the Wikipedias of the 104 largest Wikipedia languages. Its string id is `bert-base-multilingual-cased`.

XLM-RoBERTa the multilingual version of RoBERTa. Architecturally, it is identical to BERT; the only difference lies in the training regimen. XLM-RoBERTa was trained on 2TB of Common Crawl data, and it supports 100 languages. Its string id is `xlm-roberta-base`.

³ <https://commoncrawl.org/>

XLM-MLM-100 is a larger variant of XLM-RoBERTa with 16 instead of 12 layers. Its string id is `xlm-mlm-100-1280`.

distilbert-base-multilingual-cased is a *distilled* version of mBERT. It cuts the parameter budget and inference time by roughly 40% while retaining 97% of the tutor model’s NLU capabilities. Its string id is `distilbert-base-multilingual-cased`.

3.1 Subword tokenization

Subword tokenization is a key component in achieving good performance on morphologically rich languages. Out of the 5 models we compare, huBERT, mBERT and DistilBERT use the WordPiece algorithm (Schuster and Nakajima, 2012), XLM-RoBERTa and XLM-MLM-100 use the SentencePiece algorithm (Kudo and Richardson, 2018). The two types of tokenizers are algorithmically very similar, the differences between the tokenizers are mainly dependent on the vocabulary size per language. The multilingual models consist of about 100 languages, and the vocabularies per language are (not linearly) proportional to the amount of training data available per language. Since huBERT is trained on monolingual data, it can retain less frequent subwords in its vocabulary, while mBERT, RoBERTa and MLM-100, being multilingual models, have token information from many languages, so we anticipate that huBERT is more faithful to Hungarian morphology. DistilBERT uses the tokenizer of mBERT, thus it is not included in this subsection.

	huBERT	mBERT	RoBERTa	MLM-100	emtsv
Languages	1	104	100	100	1
Vocabulary size	32k	120k	250k	200k	–
Entropy of first WP	8.99	6.64	6.33	7.56	8.26
Entropy of last WP	6.82	6.38	5.60	6.89	5.14
More than one WP	94.9%	96.9%	96.5%	97.0%	95.8%
Length in WP	2.8±1.4	3.9±1.8	3.2±1.4	3.5±1.5	3.1±1.1
Length of first WP	4.3±3.0	2.7±1.9	3.5±2.7	3.1±2.0	5.2±2.4
Length of last WP	3.8±2.9	2.6±1.8	3.1±2.2	2.8±1.8	2.7±1.7
Accuracy to emtsv	0.16	0.05	0.14	0.08	1.00
Accuracy to emtsv in first WP	0.41	0.26	0.44	0.33	1.00
Accuracy to emtsv in last WP	0.43	0.41	0.47	0.39	1.00

Table 2. Measures on the train data of the POS tasks. The length of first and last WP is calculated in characters, while the word length is calculated in WPs. DistilBERT data is identical to mBERT.

As shown in Table 2, there is a gap between the Hungarian and multilingual models in almost every measure. mBERT’s shared vocabulary consists only of 120k subwords for all 100 languages while huBERT’s vocabulary contains 32k

items and is uniquely for Hungarian. Given the very limited inventory of mBERT, only the most frequent Hungarian words are represented as a single token, while longer Hungarian words are segmented, often very poorly. The average number of subwords a word is tokenized into is 2.77 in the case of huBERT, while all the other models have significantly higher mean length. This does not pose a problem in itself, since the tokenizers work with a given dictionary size and frequent words need not to be segmented into subwords. But in case of words with rarer subwords, the limits of smaller monolingual vocabulary can be observed, as shown in the following example: *szállítójárművekkel* ‘with transport vehicles’; *szállító-jármű-vek-kel* ‘transport-vehicle-PL-INS’ for huBERT, *sz-ál-lí-tó-já-rm-ű-vek-kel* for mBERT, which found the affixes correctly (since affixes are high in frequency), but have not found the root ‘transport vehicle’.

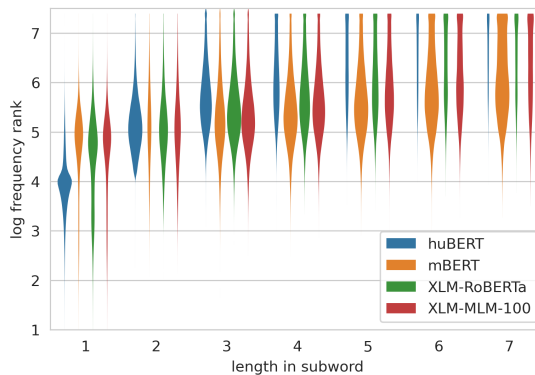


Fig. 2: Distribution of length in subword vs. log frequency rank. The count of words for one subword length is proportional to the size of the respective violin.

Distributionally, huBERT shows a stronger Zipfian distribution than any other model, as shown in Figure 2. Frequency and subword length are in a linear relationship for the huBERT model, while in case of the other models, the subword lengths does not seem to be correlated the log frequency rank. The area of the violins also show that words typically consist of more than 3 subwords for the multilingual models, contrary to the huBERT, which segments the words typically into one or two subwords.

4 Results

We find that huBERT outperforms all models in all tasks, often with a large margin, particularly in the lower layers. As for the choice of subword pooling (first or last) and the choice of layer, we note some trends in the following subsections.

4.1 Morphology

The last subword is always better than the first subword except for a few cases for degree ADJ. This is not surprising because superlative is marked with a circumfix and it is differentiated from comparative by a prefix. The rest of the results in this subsection all use the last subword.

huBERT is better than all models, especially in the lower layers in morphological tasks, as shown in Figure 3. However, this tendency starts at the second layer and the first layer does not usually outperform the other models. In some morphological tasks huBERT systematically outperforms the other models: these are mostly the simpler noun and adjective-based probes. In possessor tasks (tagged [psor] in Figure 3) XLM models are comparable to huBERT, while mBERT and distil-mBERT generally perform worse than huBERT. In verb tasks XLM-RoBERTa achieves similar accuracy to huBERT in the higher layers, while in the lower layers, huBERT tends to have a higher accuracy.

HuBERT is also better than all models in all tasks when we use the weighted average of all layers as illustrated by Figure 4. The only exceptions are adjective degrees and the possessor tasks. A possible explanation for the surprising effectiveness of XLM-MLM-100 is its higher layer count.

4.2 POS tagging

Figure 5 shows the accuracy of different models on the gold-standard Szeged UD and on the silver-standard data created with emtsv.

Last subword pooling always performs better than first subword pooling. As in the morphology tasks, the XLM models perform only a bit worse than huBERT. mBERT is very close in performance to huBERT, unlike in the morphological tasks, while distil-mBERT performs the worst, possibly due to its far lower parameter count.

We next examine the behavior of the layers by relative position.⁴ The embedding layer is a static mapping of subwords to an embedding space with a simple positional encoding added. Contextual information is not available until the first layer. The highest layer is generally used as the input for downstream tasks. We also plot the performance of the middle layer. As Figure 6 shows, the embedding layer is the worst for each model and, somewhat surprisingly, adding one contextual layer only leads to a small improvement. The middle layer is actually better than the highest layer which confirms the findings of Tenney et al. (2019a) that BERT rediscovers the NLP pipeline along its layers, where POS tagging is a mid-level task. As for the choice of subword, the last one is generally better, but the gap shrinks as we go higher in layers.

4.3 Named entity recognition

In the NER task (Figure 7), all of the models perform very similarly in the higher layers, except for distil-mBERT which has nearly 3 times the error of

⁴ We only do this on the smaller Szeged dataset due to resource limitations.

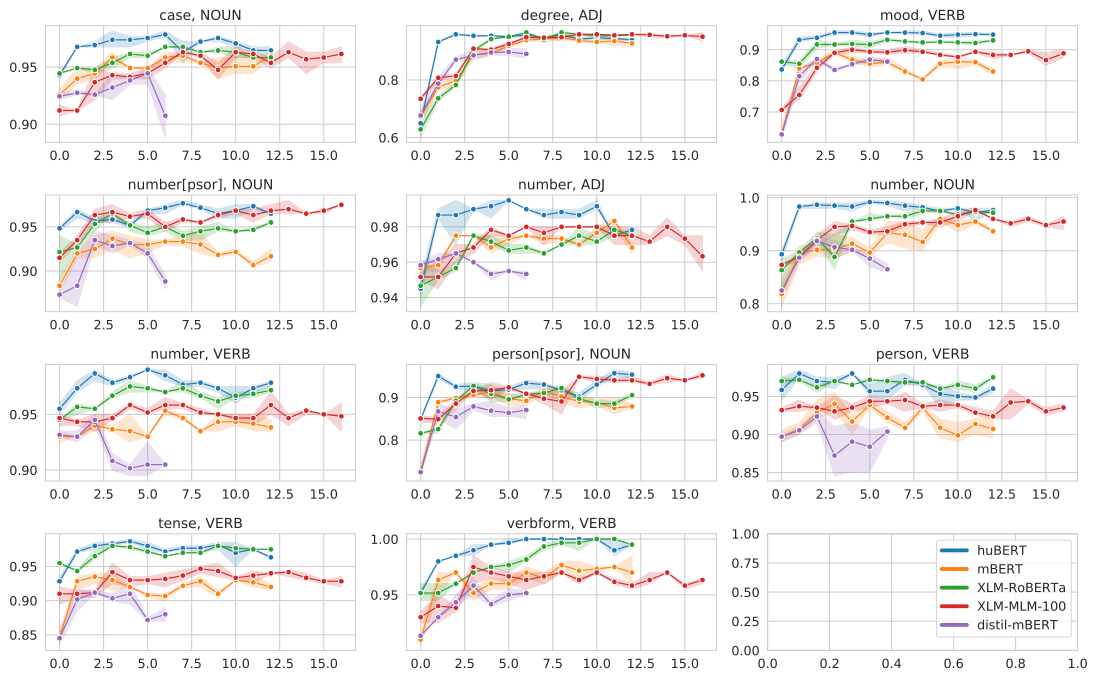


Fig. 3: The layerwise accuracy of morphological probes using the last subword. Shaded areas represent confidence intervals over 3 runs.

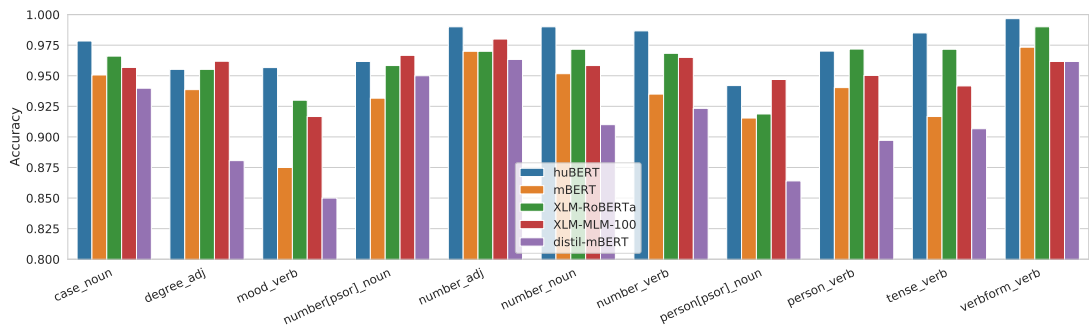


Fig. 4: Probing accuracy using the weighted sum of all layers.

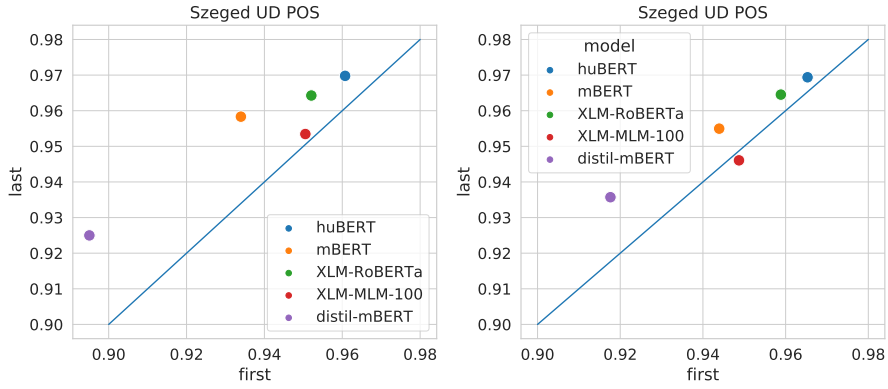


Fig. 5: POS tag accuracy on Szeged UD and on the Webcorpus 2.0 sample

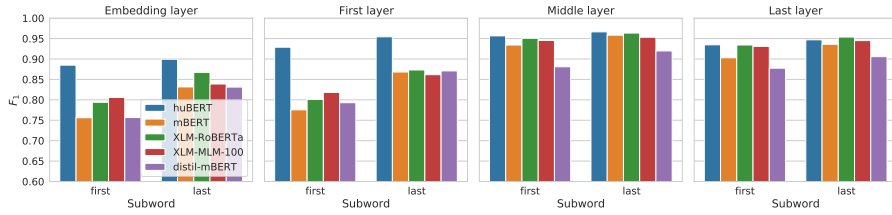
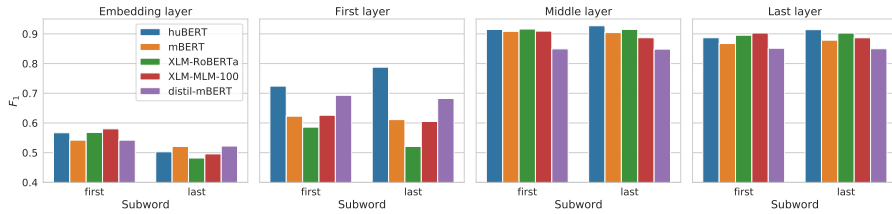


Fig. 6: Szeged POS at 4 layers: embedding layer, first Transformer layer, middle layer, and highest layer.

Fig. 7: NER F_1 score at the lowest, middle and highest layers.

the best model, huBERT. The closer we get to the global optimum, the clearer huBERT’s superiority becomes. Far away from the optimum, when we use only the embedding layer, first subword is better than last, but the closer we get to the optimum (middle and last layer), the clearer the superiority of the last subword choice becomes.

5 Related work

Probing is a popular method for exploring blackbox models. Shi et al. (2016) was perhaps the first one to apply probing classifiers to probe the syntactic knowledge of neural machine translation models. Belinkov et al. (2017) probed NMT models for morphology. This work was followed by a large number of similar probing papers (Belinkov et al., 2017; Adi et al., 2017; Hewitt and Manning, 2019; Liu et al., 2019a; Tenney et al., 2019b; Warstadt et al., 2019; Conneau et al., 2018a; Hupkes and Zuidema, 2018). Despite the popularity of probing classifiers, they have theoretical limitations as knowledge extractors (Voita and Titov, 2020), and low quality of silver data can also limit applicability of important probing techniques such as canonical correlation analysis (Singh et al., 2019),

Multilingual BERT has been applied to a variety of multilingual tasks such as dependency parsing (Kondratyuk and Straka, 2019) or constituency parsing (Kitaev et al. (2019)). mBERT’s multilingual capabilities have been explored for NER, POS and dependency parsing in dozens of language by Wu and Dredze (2019) and Wu and Dredze (2020). The surprisingly effective multilinguality of mBERT was further explored by Dufter and Schütze (2020).

6 Conclusion

We presented a comparison of contextualized language models for Hungarian. We evaluated huBERT against 4 multilingual models across three tasks, morphological probing, POS tagging and NER. We found that huBERT is almost always better at all tasks, especially in the layers where the optima are reached. We also found that the subword tokenizer of huBERT matches Hungarian morphological segmentation much more faithfully than those of the multilingual models. We also show that the choice of subword also matters. The last subword is much better for all three kinds of tasks, except for cases where discontinuous morphology is involved, as in circumfixes and infixed plural possessives (Antal, 1963; Mel’cuk, 1972). Our data, code and the full result tables are available at https://github.com/juditacs/hubert_eval.

Acknowledgements

This work was partially supported by National Research, Development and Innovation Office (NKFIH) grant #120145: “*Deep Learning of Morphological Structure*”, by National Excellence Programme 2018-1.2.1-NKP-00008: “*Exploring the*

Mathematical Foundations of Artificial Intelligence”, and by the Ministry of Innovation and the National Research, Development and Innovation Office within the framework of the Artificial Intelligence National Laboratory Programme. Lévai was supported by the NRDI Forefront Research Excellence Program KKP_20 Nr. 133921 and the Hungarian National Excellence Grant 2018-1.2.1-NKP-00008.

Bibliography

- Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., Goldberg, Y.: Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In: Proceedings of International Conference on Learning Representations (2017)
- Antal, L.: The possessive form of the Hungarian noun. *Linguistics* 3, 50–61 (1963)
- Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., Glass, J.: What do neural machine translation models learn about morphology? In: Proc. of ACL (2017), <https://www.aclweb.org/anthology/P17-1080>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2019)
- Conneau, A., Kruszewski, G., Lample, G., Barrault, L., Baroni, M.: What you can cram into a single $\text{\$!}\#^*$ vector: Probing sentence embeddings for linguistic properties. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2126–2136. Association for Computational Linguistics (2018a), <http://aclweb.org/anthology/P18-1198>
- Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S.R., Schwenk, H., Stoyanov, V.: Xnli: Evaluating cross-lingual sentence representations. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2018b)
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (6 2019), <https://www.aclweb.org/anthology/N19-1423>
- Dufter, P., Schütze, H.: Identifying elements essential for BERT’s multilinguality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4423–4437. Association for Computational Linguistics, Online (11 2020), <https://www.aclweb.org/anthology/2020.emnlp-main.358>
- Farkas, R., Vincze, V., Schmid, H.: Dependency parsing of Hungarian: Baseline results and challenges. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 55–65. EACL ’12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), <http://dl.acm.org/citation.cfm?id=2380816.2380826>

- Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4129–4138 (2019)
- Hupkes, D., Zuidema, W.: Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. In: Proc. of IJCAI (2018), <https://doi.org/10.24963/ijcai.2018/796>
- Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: emtsv – Egy formátum mind felett [emtsv – One format to rule them all]. In: Berend, G., Gosztolya, G., Vincze, V. (eds.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport (2019)
- Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014), <https://arxiv.org/abs/1412.6980>
- Kitaev, N., Cao, S., Klein, D.: Multilingual constituency parsing with self-attention and pre-training. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3499–3505. Association for Computational Linguistics, Florence, Italy (7 2019), <https://www.aclweb.org/anthology/P19-1340>
- Kondratyuk, D., Straka, M.: 75 languages, 1 model: Parsing universal dependencies universally. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2779–2795. Association for Computational Linguistics, Hong Kong, China (11 2019), <https://www.aclweb.org/anthology/D19-1279>
- Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (11 2018), <https://www.aclweb.org/anthology/D18-2012>
- Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., Smith, N.A.: Linguistic knowledge and transferability of contextual representations pp. 1073–1094 (2019a), <https://www.aclweb.org/anthology/N19-1112>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach (2019b)
- Mel'cuk, I.A.: On the possessive forms of the Hungarian noun. In: Kiefer, F., Rouwet, N. (eds.) Generative grammar in Europe, pp. 315–332. Reidel, Dordrecht (1972)
- Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)
- Nivre, J., Abrams, M., Agić, Ž., et al.: Universal Dependencies 2.3 (2018), <http://hdl.handle.net/11234/1-2895>, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

- Schuster, M., Nakajima, K.: Japanese and Korean voice search. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5149–5152. IEEE (2012)
- Shi, X., Padhi, I., Knight, K.: Does string-based neural MT learn source syntax? In: Proc. of EMNLP (2016), <https://www.aclweb.org/anthology/D16-1159>
- Singh, J., McCann, B., Socher, R., Xiong, C.: BERT is not an interlingua and the bias of tokenization. In: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). pp. 47–55. Association for Computational Linguistics, Hong Kong, China (11 2019), <https://www.aclweb.org/anthology/D19-6106>
- Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4593–4601. Association for Computational Linguistics, Florence, Italy (7 2019a), <https://www.aclweb.org/anthology/P19-1452>
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R.T., Kim, N., Durme, B.V., Bowman, S., Das, D., Pavlick, E.: What do you learn from context? Probing for sentence structure in contextualized word representations. In: Proc. of ICLR (2019b), <https://openreview.net/forum?id=SJzSgnRcKX>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- Voita, E., Titov, I.: Information-theoretic probing with minimum description length. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 183–196. Association for Computational Linguistics, Online (11 2020), <https://www.aclweb.org/anthology/2020.emnlp-main.14>
- Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.F., Phang, J., Mohananey, A., Htut, P.M., Jeretic, P., Bowman, S.R.: Investigating BERT’s knowledge of language: Five analysis methods with NPIs. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2877–2887. Association for Computational Linguistics, Hong Kong, China (11 2019), <https://www.aclweb.org/anthology/D19-1286>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), <https://www.aclweb.org/anthology/2020.emnlp-demos.6>

- Wu, S., Dredze, M.: Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 833–844. Association for Computational Linguistics, Hong Kong, China (11 2019), <https://www.aclweb.org/anthology/D19-1077>
- Wu, S., Dredze, M.: Are all languages created equal in multilingual BERT? In: Proceedings of the 5th Workshop on Representation Learning for NLP. pp. 120–130. Association for Computational Linguistics, Online (7 2020), <https://www.aclweb.org/anthology/2020.repl4nlp-1.16>

HILBERT, magyar nyelvű BERT-large modell tanítása felhő környezetben

Feldmann Ádám¹, Hajdu Róbert¹, Indig Balázs², Sass Bálint², Makrai Márton²,
Mittelholz Iván², Halász Dávid², Yang Zijian Győző², Váradi Tamás²

¹ Pécsi Tudományegyetem, Általános Orvostudományi Kar, Magatartástudományi Intézet,
Alkalmazott Adattudomány és Mesterséges Intelligencia Csoport,
7624 Pécs, Szigeti u 12.

{feldmann.adam,hajdu.robert}@pte.hu

² Nyelvtudományi Intézet,

1394 Budapest, Pf. 360

{indig.balazs,sass.balint,makrai.marton,mittelholz.ivan,
halasz.david,yang.zijian.gyozo,varadi.tamas@nytud.hu}

Kivonat A dolgozatban bemutatjuk a magyar nyelvű BERT-large modell készítését, amely 3.667 milliárd szavas szövegtudományi korpusz felhasználásával jött létre olyan megoldásokat alkalmazva, amelyek eddig egyedül angol nyelvi modellek létrehozásánál jelentek meg. A célunk olyan felhő alapú komplex számítási környezet létrehozása volt, amelyben mind szoftveres, mind pedig hardveres eszközök állnak rendelkezésre azért, hogy az új, mélytanulás alapú nyelvi modellek magyar nyelvi korpuszokkal tanítva is elérhetővé váljanak, hasonlóan a nagyobb nyelveken már elérhető state-of-the-art modellekhez. A környezet az ONNX keresztplatform megoldásait felhasználva sokkal erőforrás-optimalizáltabban hajtja végre a modellek tanítását. HILBERT, a magyar nyelvű BERT-large nyelvi keretrendszer ONNX, PyTorch, Tensorflow formátumokban rendelkezésre áll.

Kulcsszavak: BERT-large, ONNX, HILBERT, NER, Transformers

1 Bevezetés

Ebben a cikkben bemutatjuk a BERT-large nyelvi keretrendszer magyar adaptációját, az ahhoz szükséges számítási háttérrel és magát a folyamatot. A BERT-modellt (Bidirectional Encoder Representations from Transformers), amely általános célú nyelvmegértő modell, a Google AI Language kutatócsoportja 2018 októberében publikálta (Devlin és mtsai, 2018). Céljuk egy általános, komplex és kontextus érzékeny beágyazott nyelvi eszköz létrehozása volt. A modell egyedinek számított a 340 millió paraméterével, mivel ezt megelőzően a mélytanuló modellek, területtől függetlenül sokkal kisebb paraméterszámmal jelentek meg. A BERT eszköz a nyelvi megértést célzó modellek rendkívüli mennyiségű tanítóadat igényét igyekszik mederbe terelni transzfer tanulás segítségével.

A BERT-modell alapkoncepciója szerint a felhasználónak elég egy előre megtanított modellt előkészítenie, majd ezt jóval kisebb adathalmazon transzfer tanulás

segítségével adott célfeladatokhoz kell finomhangolnia. A BERT-large modell előtanítása rendkívül számításigényes feladat, különleges technológiai háttérrel igényel ennek megvalósítása, amely hatvannégy darab V100-as NVIDIA GPU felhasználásával közel 100 óra futásidőt vesz igénybe.

A BERT modelleknek két fajtája érhető el méretük szerint; az első a BERT-base, amely 110 millió paraméterrel rendelkezik, illetve a BERT-large, amely 340 millió paramétert tartalmaz. Devlin eredeti célja a BERT-base megalkotásával az volt, hogy a BERT modellt összevethesse a korábban megjelent, szintén 110 millió paraméteres GPT (Generative Pretrained Transformer) névre hallgató eszközzel. Mindkét BERT modell azonos architektúrával rendelkezik, de paramétereikben különböznek. A BERT-base 12 darab kódoló réteggel, míg a BERT-large modell 24 darab kódoló réteggel bír. További különbség, hogy a kódoló rétegen belül nagyobb a figyelmi fej; a BERT-base 12, míg a BERT-large 16 figyelmi fejjel rendelkezik. A feedforward rétegen belül, mely a kódoló réteg egyik része, 768 rejtett feldolgozó elem található a kisebb, míg 1024 a nagyobb modellnél.

A BERT-base változatot magyar nyelvre Nemeskey Dávid készítette el (Nemeskey, 2020), demonstrálva a modell kiemelkedő képességeit különböző nyelvi feladatokon. A BERT modell részletes, kellő mélységgel történő tárgyalása szintén Nemeskey Dávid előbb hivatkozott publikációjában olvasható. Jelen tanulmány a szükséges számítási környezet jellemzésére és bemutatására helyezi a hangsúlyt, valamint a BERT-large modell előtanítását és finomhangolását mutatja be.

2 A HILBERT modell létrehozása

2.1 Számítási környezet kialakítása

Az extrém nagy méretű mélytanuló modellek tanításhoz speciális hardver és szoftver-környezet szükséges. Mivel a GPU alapú számítási eszközök közül is csak a kifejezetten gépi tanulás támogatására létrehozott célprocesszorok alkalmasak, valamint ezekből több darabra is szükség van a tanításhoz, a felhő alapú számítási megoldások felé fordultunk. A Microsoft Azure felhőszolgáltatáson belül találtunk megfelelő méretű, bérelhető számítási kapacitást és szoftveres környezetet. Az AzureML környezetet kifejezetten gépi tanulási folyamatok megvalósítására és szolgáltatására fejlesztették. Modulokra bonthatóan kezelhetőek benne az egyes részfeladatok, melyhez tárolókat és egyéb erőforrásokat rendeltünk. Az AzureML SDK 1.6-os változatát használtuk Python 3.6 nyelven. A mélytanulási feladathoz pedig a PyTorch framework-öt választottuk az ONNX Runtime keresztplatform felhasználásával. A PyTorch szabványosan elérhető AzureML környezetben, az ONNX platform pedig integrálja a legújabb számításoptimalizáló és gyorsító megoldásokat, köztük a DeepSpeed technológiát (Rajbhandari és társai, 2019), amely akár ötszörösére gyorsítja a modellek tanítását a GPU memória használatának optimalizálásán keresztül. A szükséges számítási klasztert is itt hozzuk létre, ahol az AzureVM eszközök közül választhatjuk ki a feladathoz leginkább megfelelő tulajdonságokkal bíró csomópontokat. Kezdeti lépésként létrehoztunk egy eszközcsoportot az Azure-ben, melyben számítási csomók és tárolók egyaránt helyet kaptak. Fontos, hogy nagyobb adatmozgás esetén a virtualizált környezet ellenére az egyes

eszközök fizikailag is közel legyenek egymáshoz, mert a tárhelyműveletek sokmilliószor lassabbak, mint a számítási műveletek. Mivel a modell tanításához GPU alapú erőforrás szükséges, de a kód szerkesztése, módosítása ezt nem kívánta meg, így létrehozunk egy alapértelmezett számítási eszközt egy virtuális gép segítségével. Az allokált eszköz elegendő a környezet felparaméterezéséhez és a tárolókkal történő műveletek végrehajtásához.

2.2 Az adatok jellemzése

Az előtanító korpusz

A nyelvi modellek készítésének döntő fontosságú kérdése a korpusz minősége, amelyen a modell előtanítása készül. Az előtanításhoz szükséges korpuszt a nyelvmodell célja szabja meg. A mai gyakorlatban az honosodott meg, hogy rendszerint egy *általános célú* nyelvi modellt készítenek, melyet aztán adott feladat számára *finomhangolnak*. Az általános célú nyelvmodellt olyan korpuszon célszerű betanítani, amely a nyelvhasználat széles körét reprezentálja. A nyelvhasználat egészét átfogóan és arányaiban is modellálni nem jól definiált feladat, mert szigorú értelemben vett reprezentatív mintát nem lehetséges összeállítani. Ugyanis a teljes populációról (azaz a nyelvhasználat egészéről) nincsenek megbízható adataink. A legtöbb, amit tehetünk az, hogy egy úgynevezett kiegyensúlyozott korpusz (balanced corpus) összeállítására törekszünk, illetve figyelembe vesszük a korpusz felhasználásának a célját.

A BERT modellhez szükséges legalább 3,5 milliárd szónyi folyó szövegből álló korpuszt az alábbi forrásokból állítottuk össze.

MNSZ. Fontos forrás a Nyelvtudományi Intézetben készült Magyar Nemzeti Szövegtár. Egyrészt hat stílusrétegből (sajtó, szépirodalom, tudományos, hivatalos, személyes, beszélnyelvi) tartalmaz szövegeket, másrészt ezen belül öt regionális nyelvváltozatra oszlik. A regionális nyelvváltozatok az egyes határon túli magyar területeket képviselik. Kiemelendő az önmagában is jelentős, 76 millió szavas beszélnyelvi (rádiós) alkorpusz, ez az MR1 Kossuth rádió bizonyos anyagainak öleli fel az 2004-2012 évekből, felolvasott szöveget (hírek) és spontán beszélgetést (riportok) vegyesen. Mérete 975 millió szó.

JSI. A szlovén Jožef Stefan Institute az eventregistry.org címen futó webszolgáltatás céljaira 2013 óta számos nyelven gyűjti a híreket internetes forrásokból (RSS-ből). Ennek a magyar anyagát használtuk fel. Ebben egészen friss hírek is szerepelnek, megjelennek az aktuális témák (koronavírus stb.). Mérete 1,06 milliárd szó.

NOL. A MNSZ sajtókorpuszát kiegészítettük a Mediaworkstól kapott *Népszabadság online* anyaggal. Ennek terjedelme 48 millió szó.

OS. A következő forrás a szabadon hozzáférhető filmfelirat-adatbázis, az opensubtitles.org magyar része. Amint említettük, erre jellemző a beszélnyelvi stílus, rövid mondatok, párbeszédes forma. Mérete 471 millió szó.

KM. Az utolsó forrás egy jelentős, nyilvános közösségi média posztokból és kommentekből származó szöveganyag, melyet a Neticle Kft-től kaptunk meg korábban. Mérete 1,11 milliárd szó.

A szótár

Több milliárd szavas korpusz esetén a rendszer által használt szótár kritikus jelentőséget kap. A kihívást az jelenti, hogy a szótárnak lehetőleg le kell fednie a korpuszban előforduló szóalakok egészét, ugyanakkor kis méretűnek kell lennie a hatékonyság jegyében. A szavak belső reprezentációjára egy olyan szótárt használ, amelyekben a szavak statisztikai alapon *szóelemekre* vannak bontva, extrém esetben az egyes karakterekig. A BERT modell a Google által kifejlesztett WordPiece eljárást alkalmazza. A szótárak mérete általában 30 és 50 ezer elem között váltakozik. A magyar nyelv morfológiai sajátosságaira tekintettel a HILBERT modellhez 64000 elemes WordPiece szótárt fejlesztettünk ki. A szótár hatékonyságát Nemeskey Dávid kódjával mértük. Minél kevesebb szóelemre bontja a szótár a felszíni szavakat, annál jobbnak mondható. A HILBERT tanításánál használt WordPiece esetében ez a mutató 1, 15, azaz átlag egy szövegszót 1,15 szóelemre bont a tokenizáló.

2.3 Az adatok előfeldolgozása

A modell tanításához elsőként a szövegeket bináris formába kell hozni ahhoz, hogy a BERT modell tanításához felhasználhatóak legyenek. Az eredeti BERT modellek a Wikipédia angol nyelvű szövegtörzsén és könyvtörzsön készültek. A magyar szövegek előkészítése során meghagytuk az eredeti, Wikipédiára utaló könyvtörzskezelést. A nyers szövegfájlok összmérete 25 GB. A szöveg darabolására az előfeldolgozási lépések memóriaigénye miatt volt szükség. Az előfeldolgozás egy külön folyamat, melynek bemenete a 100 darab szövegfájl és a kimenete olyan bináris állomány, amelyben a tenzor bemenetek vannak elrendezve modelltanításhoz és validációhoz. Az adatfeldolgozáshoz külön programot készítettünk. A szöveg rendezése során a beolvasott szöveget úgy tisztítjuk, hogy csak az alfanumerikus és központosító karakterek maradjanak benne, illetve minden sorba egy mondat kerüljön. Ezután speciális tokeneket kell hozzáadni a tokenizált szöveghez <cls> és <sep> elválasztó karaktereket. A <cls> a szövegek különböző osztályozásakor játszik szerepet, míg a <sep> szeparátorként választ el mondatokat egymástól. A program iteratívan végighaladva az aktuális szövegrészen illeszti hozzá a szótárban található szóelemeket, ahol nem ismert szóelem tokenel találkozunk, ott azt <unk> taggal helyettesíti. Ez a folyamat többféle szótárral, illetve tokenizáló eszközzel is történhet. A tokenizálás 25 GB szövegen 4 nap alatt futott le. A folyamat memória intenzív feldolgozás, ahol a számítások végrehajtásához egy STAN-DARD_D14_V2 virtuális gépet vettünk igénybe. Ennek eredményeként egy blob tárolóban létrejött 100 db bináris állomány 600GB körüli tárhely igényel, mely már készen állt a BERT-large modell tanításához.

2.4 GPU klaszter létrehozása

A modell számítási paramétereinek a megállapítása szorosan összefügg a felhasználható, rendelkezésre álló GPU kapacitás méretétől. Mivel GPU segítségével nagyon gyorsan lehet mátrixokat összeszorozni és feldolgozni, ezért kiválóan alkalmasak tenzor alapú számítások futtatásához, sokszoros teljesítménynövekedést nyújtva a CPU alapú feldolgozással szemben. A leginkább elterjedt eszközök az NVIDIA által gyártott

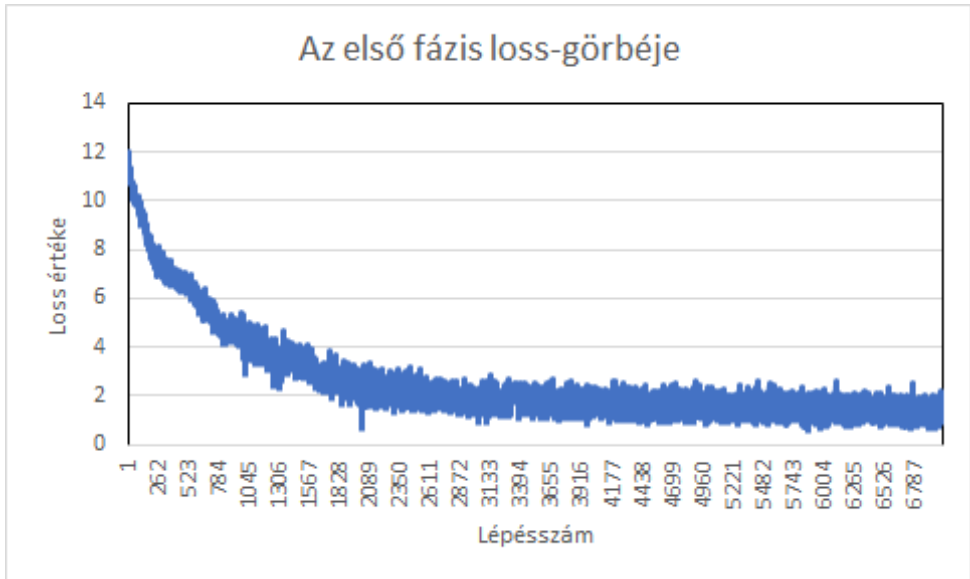
V100-as GPU-k, melyekhez különböző méretű VRAM tartozik. Az Azure környezetben elérhető, GPU alapú számítási csomópontok közül az NCv3-as széria NC24rs v3 kódjelű node-jára esett a választásunk. Ez az eszköz 4 db V100-as GPU-t tartalmaz, melyekhez egyenként 16GB VRAM tartozik a 448GB RAM mellett. Azért választottuk ezt a számítási csomót, mert RDMA-kompatibilisek és Infiniband alapú kapcsolat segítségével rövid látencia mellett biztosítják a számítási fűtőn belül a node-ok közötti, alacsony szintű kommunikációt. Ez azért különösen fontos, mert MNI (Message Passing Interface) segítségével jobban párhuzamosíthatóak a több GPU-s feldolgozást igénylő feladatok, ha több csomópontot szeretnénk összekötni.

2.5 A tanítási paraméterek megadása

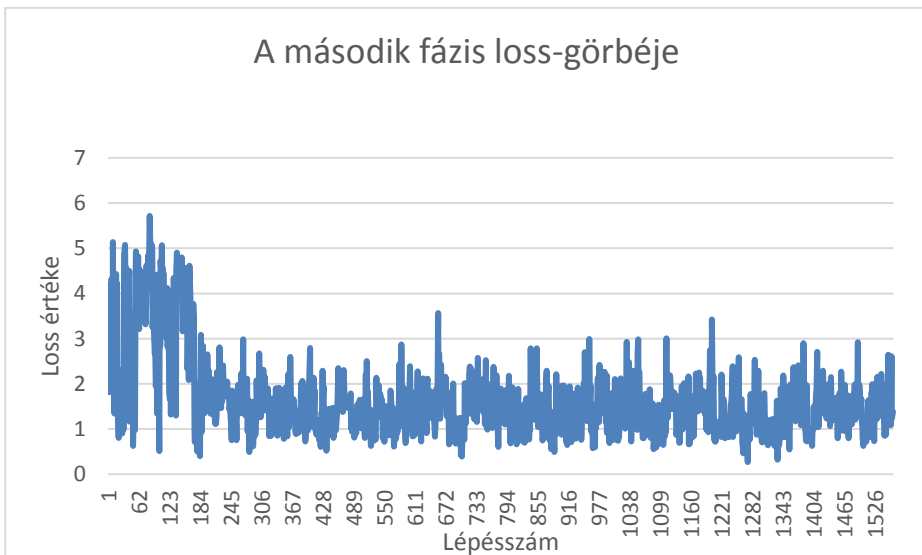
A BERT-large modell tanításához az AzureML Kísérlet modulján belül kell konfigurálni az MPI-t és meg kell adni, hogy egy számítási csomóban hány darab GPU található. Meg kell adni továbbá, hogy a GPU-ban található CUDA magok kezeléséhez szükséges csomagokat és az openmpi drivereit melyik docker image tartalmazza. A batch size paraméter függ a rendelkezésre álló GPU-k számától, illetve azok VRAM méretétől. A párhuzamos GPU használat esetén minden GPU külön számol grádiens loszt különböző adatokon. Minél nagyobb a grádiens mérete, annál inkább csökken a zaj hatása a tanításra. Ennek ellenére a tanítás későbbi szakaszában a nagy grádiens méret kevésbé vezet optimális eredményhez.

A modell tanítását az NVIDIA scriptjével végeztük, amely két fázisra osztja a tanítást. Az első fázisban 128 token hosszúságú modellt készítünk, majd ezt követően 512 token hosszúsággal folytatjuk tovább a modell tanítását. Erre bontásra azért van szükség, mert a figyelmi fejek méretének növekedésével a számítási kapacitás négyzetesen növekszik. A második fázis gyakorlatilag egy finomhangolási lépés. A modell előtanításának 90%-a 128-as hosszúsággal, míg az utolsó 10% 512-es tokenhosszra történik (Devlin és mtsai, 2018). Az első fázis 7038 lépést, míg a második 1563 lépést tartalmazott.

A szkript paramétereit a kötegméret, a grádiens akkumuláció és a GPU memória limitet kivételével az alapértelmezett értékeken hagytunk. A modell 128-as szekvencia hosszon 32-es batch mérettel, $6e-3$ tanulási rátával (0.2843 előmelegítési ráta), 512-es szekvencia hosszon pedig 8-as batch mérettel, $4e-3$ tanulási rátával (0.128 előmelegítési ráta) paraméterekkel tanult. Ezeket mindkét fázis esetében külön meghatároztuk a használt számítási csomókhoz. A modell tanítási folyamatának állapotáról a tanulás veszteség-függvénye nyújt információt (Ábra1, Ábra2). A magyar BERT-large (HIL-BERT) tanítása során támpontként szolgált az NVIDIA által közzétett veszteség görbe az angol nyelvi modellhez, illetve ugyanezeket megkaptuk a Microsoft fejlesztői csapatától is.



1. ábra A BERT-large modell veszteségfüggvénye tanítás során 128 token hosszúságú szekvenciákkal.



2. ábra A BERT-large modell veszteségfüggvénye tanítás során 512 token hosszúságú szekvenciákkal.

2.6 A kész modell finomhangolása

Az elkészített modell finomhangolása egy gyakran használt transzfer tanulási módszer. Ezzel a felügyelt tanítási módszerrel specifikus feladatokra lehet tovább tanítani a modellt, mint a névelemek felismerése, vagy kontextusalapú kérdés-válasz generálásra, illetve különféle célú szövegosztályozó feladatok végrehajtása. Működését tekintve az előtanított modell utolsó rétege fölé egy klasszifikációs réteg kerül (Devlin és mtsai, 2018), ami a tovább tanítás során a bemenetet és annotációit tanulja meg.

A névelem-felismerés egy gyakran használt módszer a nyelvi modellek teljesítménymérésére. A szegedi Corpus of Business Newswire Texts (szegedNER) corpust alkalmaztuk a névelem-felismerés tanításához (Szarvas és mtsai, 2006). A korpuszt 80-10-10 arányban bontottuk fel tanító, validációs és teszt adathalmazokra. A transzfer tanítási megoldás a transformers könyvtár példái közül lett kiválasztva. Az F1-értékek számítása a sequeval könyvtárral történt. A finomhangolás feladathoz NVIDIA Tesla V100 16GB videokártyát használtunk felhőkörnyezetben. A finomhangolási paraméterek közül, a modell 3e-05-ös (lineárisan csökkenő) tanulási rátán és 3 epoch-on keresztül tanult, 8-as kötegmérettel.

A modell validációs F1-értéke a corpusban annotált 16 névelemosztályra összesen 95.39%-ot adott. A modell valódi képességeit leginkább jelző, a teszt adathalmaz F1-értékére, szintén 16 névelem osztályra 93.91%-ot kaptunk. Ezek az értékek azt mutatják, hogy a magyar nyelvű HILBERT teljesítménye a névelem keresés terén rendelkezik a BERT-large modellektől elvárt képességekkel (Virtanen és mtsai 2019; Martin és mtsai, 2019).

A többi modellel való összevetés lehetőségét árnyalja, hogy a nemzetközi szakirodalomban az egyes, különböző nyelveken elérhető, annotált névelem adatbázisok sokszor eltérő névelem kategóriákat(is) tartalmaznak, illetve az elérhető adatbázisok mérete nyelvenként nagyon eltérő lehet. További nehézség, hogy az egyes modellek tanítása gyakran eltérő epoch-számmal történik. Ezeken túl a finomhangolás random inicializálása is hatással van a finomhangolt modellek teljesítményére (Dodge és mtsai, 2020).

3. Összegzés

A HILBERT, magyar nyelvű BERT-large modellt sikerült létrehoznunk egy kereskedelmi számítási felhőben, ahol olyan horizontálisan és vertikálisan is skálázható infrastruktúrát alakítottunk ki, amelyben több, akár magasabb paraméterszámú modellek előállítására is lehetségessé vált. Elkészítettük a szegedNER corpus (Szarvas és mtsai, 2006) segítségével a modellünk finomhangolását névelemkereséshez, amelyben ~94%-os teszt eredményt sikerült elérnünk. Jelenleg is rendelkezésre áll több, a modellhez köthető alkalmazásunk, amelyben a HILBERT, mint extraktív szövegösszegző, illetve mint keresőmotor jelenik meg. A BERT-large igazi előnye azonban a kérdés-válasz típusú feladatokban mutatkozik meg a többi, kisebb paraméterszámú modellhez képest, de ilyen adathalmaz magyar nyelven egyelőre nem elérhető.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik ki Varga Gábornak és a Microsoft Magyarország Kft. többi munkatársának a segítségükért, akik lehetővé tették, hogy a pandémiás időszak alatti korlátozások ellenére hozzáférjünk a szükséges számítási kapacitásokhoz. Külön szeretnénk megköszönni a lehetőséget, hogy a Microsoft Corporation ONNX Runtime fejlesztőcsapatával együtt dolgozva a legújabb fejlesztéseiket tesztelve tudtuk létrehozni a magyar nyelvű BERT-large modellt.

Hivatkozások

- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deepbidirectional transformers for language understanding. In: Proc. of NAACL (2019)
- Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., Smith, N.: Fine-tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. (2020)
- Martin, L., Muller, B., Ortiz, S., Dupont, P.J. Romary, L., Villemonte de la Clergerie, E., Seddah, D., Sagot, B., CamemBERT: a Tasty French Language Model (2019)
- Nemeskey, D.M.: Egy embert próbáló feladat. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 409–418. Szegedi Tudományegyetem, Szeged (2020a)
- Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimization towards training a trillion parameter models. (2019)
- Szarvas, G., Farkas, R., Kocsor, A.: A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms. In: Discovery Science, 9th International Conference, DS 2006, Barcelona, Spain, October 8–10, 2006, Proceedings. pp. 268–278 (2006)
- Virtanen, A., Kanerva, J., Ilo, R., Louma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: BERT for Finnish (2019)
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding (2018)

TRANSZKRIPCIÓ, TRANSZLITERÁCIÓ

Magyar hadifoglyok adatainak orosz-magyar átírása és helyreállítása, és a szabadszöveges adatbázisok tulajdonságai

Sass Bálint, Mittelholcz Iván, Halász Dávid, Lipp Veronika, Kalivoda Ágnes

Nyelvtudományi Intézet, ELKH, MTA
{sass.balint,mittelholcz.ivan,lipp.veronika,kalivoda.agnes}@nytud.hu
david.peter.halasz@gmail.com

Kivonat Ebben a tanulmányban a magyar hadifoglyok adatbázisában lévő tulajdonnevek orosz-magyar átírásának módszerét és tanulságait mutatjuk be. Az adatbázisban a 682000 hadifogoly adatai cirill betűkkel leírva állnak rendelkezésre. Az adatok két körben szenvedtek torzulást: először, amikor az adatokat felvevő szovjet katona hallás utána leírta, majd mikor 60 év múltán szintén orosz anyanyelvűek manuális munkával digitalizálták az anyagot a kézzel írott kartonokról. Esetünkben nem szimpla átírásról van szó, hanem valójában az eredeti magyar szó helyreállításáról. Külön feladatot jelentett a helyeket leíró adatok adatmezőkre bontása. Szabályalapú algoritmusunkban szigorú és laza átírást, valamint közelítő keresést alkalmazunk, az átírást listákkal vetjük össze. Ha egyik módszer sem vezet eredményre, akkor a buta betűről-betűre átírást adjuk vissza. Eredmény: az adatok 77%-ához tudtunk helyes helyreállított alakat rendelni. Megfogalmazunk tanulságot a kézzel készült, korlátozatlan, szabadszöveges adatbázisok szükségszerű következtetlenségéről; valamint arról, hogy egyedi adatnál, tanulóadat híján van létjogosultsága a szabályalapú módszereknek.

Kulcsszavak: hadifogoly, átírás, transzkripció, szabályalapú, szabadszöveges adatbázis

1. Bevezetés

Harminc évnek kellett eltelnie a rendszerváltás óta, hogy Magyarország megkap hassa a II. világháború végén elhurcolt 682000 magyar hadifogoly nyilvántartási adatait. 2019-ben született meg a megállapodás a Magyar Nemzeti Levéltár és az Orosz Állami Hadilevéltár között az adatok átadásáról, és még ebben az évben meg is érkezett az anyag a Magyar Nemzeti Levéltárba. Az adatokat két formában kaptuk meg: (1) az eredeti kézzel, cirill betűkkel írt nyilvántartó kartonok digitalizált (szkennelt) változata; (2) cirill betűs leirat, adatbázis, amely a nyilvántartó kartonokon szereplő, az egyes személyekhez köthető információkat tartalmazza.

A Levéltár célja, hogy az információkat online kereshető formában közzétegye, lehetővé téve, hogy a leszármazottak hozzájussanak a rokonaikról tudható

információkhoz, illetve, hogy általánosságban kutathatóvá tegye az anyagot a szakemberek és a nagyközönség számára.

Az adatbázis jellemzője, hogy keletkezési helyének megfelelően minden adat *cirill betűkkel* szerepel benne. A kereshetőség biztosítása érdekében fontos feladat tehát az anyag átültetése magyarra. Ezt a feladatot végeztük a Nyelvtudományi Intézetben, erről számol be ez a tanulmány. A munkálatokban a Levéltár, a Helion Kft. és a Nyelvtudományi Intézet működött közre.

2. A feladat

A feladat tehát az, hogy magyarra alakítsuk az eredeti cirill betűkkel¹ leírt adatokat.

Ковач Йожеф → Kovács József

1. ábra: Alappélda. A *Kovács József* cirill formája és a helyreállítandó magyar változat. Itt elegendő, ha minden cirill betűt egyszerűen a magyar megfelelőjére írunk át.

Az 1. ábrán vázolt egyszerűnek tűnő feladat nehézségét több tényező adja. Egyrészt az a nyelvi tény, hogy az orosz betűk illetve hangok nem egy az egyben felelnek meg a magyar betűknek és hangoknak. Másrészt az anyag az elmúlt évtizedek során több alkalommal is torzult. A szovjet hadifogolytáborba érkező katonák általában nem rendelkeztek iratokkal (Katona és Szikla, 2014). Ennek megfelelően az adataikat legtöbbször *hallás után* írta le az adatokat felvevő szovjet katona. Az első torzulást tehát az okozza, hogy hallás után került rögzítésre a leíró által nem értett magyar nyelvű adat. A második torzulás akkor keletkezett, mikor az elmúlt években az Orosz Állami Hadilevéltár munkatársai manuális munkával digitalizálták a sokszor nagyon nehezen olvasható szkennelt kézirásos kartonokat, azzal együtt, hogy ők sem *értették* a leírtakat, sokszor csak a cirill betűsorozatokat igyekeztek beazonosítani és rögzíteni. A fenti két torzulás biztosra vehető, ezenkívül még lehetséges egy harmadik is: a kartonok valószínűleg nem közvetlenül a táborokban készültek, hanem egy központi helyen, így még egy másolási lépés beiktatódhat.

A történeti szövegeknél látunk hasonló jelenséget, mikor a hibázások és többszöri másolás eredményeképpen jelentősen meg tud változni az eredeti szöveg (Haader, 2014). Munkánk előzményének tekinthető Prószéky és mtsai (2002) cikke, mely a különböző forrásokból eredő karakterhibák javításával foglalkozik.

¹ A cirill betűk, az orosz ábécé és hangkészlet ismeretét feltételezzük a továbbiakban. Ld. pl. https://hu.wikipedia.org/wiki/Orosz_ábécé

3. Fordítás, átírás, helyreállítás

Ebben a részben az alapfogalmakat világítjuk meg: az elvégzendő feladat nem fordítás, nem egyszerű átírás (Bradley, 2020), hanem valójában – nevezhetjük így – helyreállítás.

orosz → magyar		
fordítás	конец	vége
átírás	конец	kányec
helyreállítás	Ковач	Kovács

1. táblázat. A helyreállítás viszonya a fordításhoz és az átíráshoz.

Ahogy az 1. táblázatban látni fogjuk, fordítás során a forrásnyelvű, forrásnyelvi írással írt szöveget tesszük át célnyelvű, célnyelvi írással írt szöveggé; átírás során ugyanezt, a forrásnyelvet megtartva célnyelvi írással átírt szöveggé; a helyreállítás során viszont az eleve célnyelvű, viszont forrásnyelvi írással írt szöveget alakítjuk a célnyelvet megtartva írásában is célnyelvivé. A lényeg tehát, hogy a szóban forgó feladatban – bár cirill betűkkel van leírva – a kiinduló elem egy *magyar* szó, ezt kell az eredeti magyar formájában és értelmében helyreállítani. Nemcsak át kell írni, hanem rá is kell jönni, hogy mi az. A feladat így jóval nehezebb, mint az egyszerű átírás, ahol a szó értelmére nem kell figyelmet fordítani.

4. A helyreállítás szintjei

Az adott nyelvi adat helyreállítása nehézségének megfelelően három szintet különítünk el.

Az első **#1** szint a betűnkénti egyértelmű átírást jelenti: 'Вилмом' → Vilmos², mikor az orosz betűket magyar megfelelőjükre cseréljük. Ez történhet a hivatalos szabályzat (Zoltán, 1981) szerint, ugyanakkor látjuk, hogy ez a módszer a legritkább esetben elegendő.

Említettük, hogy az orosz és magyar betűk sok esetben nem egy az egyben felelnek meg egymásnak. Esetünkben az jelenti a problémát, hogy *egy* orosz betűnek *több* magyar is megfelelhet (2. táblázat). A második **#2** szinten ezt a problémát oldjuk meg lényegében azáltal, hogy végigpróbálgatjuk a lehetséges betűket, hogy értelmes magyar szót kapjunk.

² Ha hangsúlyozni akarjuk, hogy egy szövegrész orosz, akkor sima idézőjelekkel, magyar esetén pedig aláhúzással jelöljük.

orosz	→ magyar
Моноки	Monoki
Миклош	Miklós
Колмар	Kalmár
Теглош	Téglás

2. táblázat. Egy orosz betűnek több magyar is megfelelhet, az orosz 'o'-nak például leggyakrabban o, ó, a vagy á.

A fentiekhez adódnak hozzá a 2. részben említett torzulások, azaz amikor adott pozícióban egyáltalán nem a megfelelő betű szerepel, illetve amikor félrehallás, félreolvasás és a szöveg nem értése miatt különféle összetettebb hibák és következtelenségek kerülnek az anyagba (3. táblázat). Ezeket a harmadik #3 szintű „okos” helyreállítás igyekszik minél jobban megoldani.

orosz	→ magyar	hibafajta
Баконьеомбандхель	Bakonyszombathely	más betű: 'c' helyett 'e' fölösleges: 'h'
Балашадарма	Balassagyarmat	kimarad: 't'
Бикшичаба	Békéscsaba	csere: 'иш' helyett 'ши'
Бешенелект	Besenyőtelek	más betűcsoport

3. táblázat. Hibafajták.

Míg a Вилмош → Vilmos esetén elegendő az #1 szint, a Ковач → Kovács megoldásához szükséges a #2 szint, mivel meg kell állapítani az orosz 'a' aktuális megfelelőjét. Végül a Шаторомойгел → Sátoraljaújhely esetén #3 szint szintű eljárás szükséges a megfelelő helyreállításhoz: a betűmegfeleltetéseken kívül fel kell oldani a 'омо' → alja konverziót is. A 4. táblázatban néhány könnyebb és nehezebb példa látható.

Nagyon valószínű, hogy az első torzulás (ld. 2. rész) során keletkeztek a félrehallásos, elírásjellegű hibák, a második torzulás során pedig a félreolvasásos, OCR-jellegű hibák. Az előbbire példa a Репцелор (4. tábl/1.), mivel a k→g félrehallás is könnyen elképzelhető, a két betű formája viszont eltérő; az utóbbira pedig a Лепец (4. tábl/7.), mert a n→t félrehallás nem valószínű, viszont a megfelelő orosz betűk ('h' és 't') alakja hasonló. Mindenesetre az anyagban a két hibaosztály szerencsétlen keveredését látjuk.

A 4. táblázat azt is illusztrálja, hogy ennél a feladatnál (és más hasonlóknál is) minden bizonnyal meglesz a táblázatban látható három osztály: a géppel megoldható esetek, a géppel nem, de manuális munkával megoldhatók és a manuális munkával sem megfejthetők. Célunk a második osztály méretének csökkentése

orosz	#1 szint	#3 szint	ember	gép
1. Репцелог	Repcelog	Répcelak	✓	✓
2. Фейняшлидке	Fejnyáslidke	Fényeslitke	✓	✓
3. Хатоваин	Hatovain	Hatvan	✓	✓
4. Лайошминш	Lajosmins	Lajosmizse	✓	✓
5. Яцберин	Jácberin	Jászberény	✓	✗
6. Ямуш	Jámus	János	✓	✗
7. Леретц	Leretc	Ferenc/Lőrinc	✓	✗
8. Блодентмигайн	Blogyentmigájn	→ 11. oldal	?	✗
9. Аирг	Airg	???	?	✗
10. Алохупкуя	Alohupkuja	???	?	✗

4. táblázat. Példák hozzávetőleges nehézségi sorrendben. Az első négy példát a jelenlegi rendszerünk jól kezeli. Az 5-7. példák emberi intelligenciával biztosan kitalálhatók, bár előfordul, hogy több megoldás is jónak tűnik.

az első növelése révén, azaz géppel minél jobban megközelíteni az emberi teljesítményt.

5. Feldolgozott adatmezők

A munkálat során azokkal az adatmezőkkel foglalkozunk, amelyeket nem fordítani, hanem helyreállítani kell (vö: 3. rész), azaz ami nem orosz szöveg, hanem magyar szöveg cirill betűkkel.

Ide tartozik: (1) vezetéknev; (2) keresztnév; (3) apai keresztnév; (4) születés helye; (5) fogságba esés helye. A nálunk megszokott „anya neve” helyett a szovjet hatóságok – az orosz nevek felépítésének megfelelően – az apai keresztnévet jegyezték fel. A két hely mező ország, megye, járás, település, utca, házszám részekből, illetve ezek épp jelenlévő részhalmazából áll. Az utcával és a házszámmal nem foglalkozunk, főképp azért, mert nem áll rendelkezésre az utcaneveket tartalmazó átfogó lista.

A fordítást igénylő mezők tehát kimaradnak: dátumok, fogadó tábor, rendfokozat, amennyiben elhunyt hol nyugszik, elbocsátó tábor, megjegyzés, azonosítók.

6. A helyreállítás módszere

A bevezetésben említett cirill betűs adatbázisból indultunk ki, a szkennelt kartonokkal nem foglalkoztunk. Utóbbi egy teljesen más léptékű feladat lenne, amit az orosz partner manuális munkával lehetőségeihez képest megfelelően elvégzett. Rendszerünkben azt a megközelítést választottuk, hogy az eredeti cirill verzióból származtatunk több lehetséges magyar átírást, Morse (2005) éppen az ellenkező irányt választotta.

6.1. Előfeldolgozás

Három részfeladatot végzünk el az előfeldolgozás keretében: a női keresztneveket és az orosz formában megjelenő apai neveket speciálisan kezeljük, valamint a helyadatokat releváns adatmezőkre bontjuk.

A foglyok között kb. 1%-ban fordulnak elő nők. Ha keresztnévként férfi és női neveket is elfogadunk, akkor számos olyan hiba adódik, hogy férfi nevet nőiként kezelünk: Пауль → Paula (helyesen: Paul/Pál); Матия → Maja (helyesen: Matija/Mátyás); Алоис → Aliz (helyesen: Alois/Alajos). Ezért azt a megoldást választottuk, hogy a keresztnevекnél csak férfi neveket fogadunk el, a női neveket pedig egyedileg kezeljük, azaz listába gyűjtve egyenként állítjuk helyre.

Az apai keresztnév mezőben sokszor megjelenik az oroszra jellemző -вич/-вна végződés, akár egyértelműen magyar névhez illesztve is: Чилик Юзеф Имревич → Csilik József Imre**vi**cs. Ezt a végződést elhagyjuk.

A hely mezőket (születés helye és fogságba esés helye) valódi adatmezőkre bontjuk: 1. ország, 2. megye, 3. járás, 4. település. A felbontást az országok listája és a meglévő rövidítések alapján végezzük (5. táblázat), a feladat a sorrendi és egyéb következtelenségek miatt okoz nehézséget.

Венгрия, обл. Пешт, д. Вечешь
→ Magyarország, Pest megye, Vecsés település
с. Сигатуйфалу, обл. Пештмеде, Венгрия
→ Magyarország, Pest megye, Szigetújfalu település

5. táblázat. Két könnyen felbontható helyleírás. Rövidítések: обл. = область → megye, д. = деревня → falu. Az esetenként megjelenő -меде tagot elhagyjuk.

6.2. Átíró szabályok

Az átírást végző szabályrendszerek létrehozásához manuális munkával megvizsgáltuk az egyes cirill karakterek 100-100 véletlenszerű előfordulását és ez alapján állapítottuk meg a lehetséges magyar megfelelőket. Az átíró szabályok két változatban készültek. A szigorú vagy strict változat egy magyar megfelelőt tartalmaz (pl. 'д' → d), a laza vagy loose pedig többet (pl. 'д' → d|gy|t). Az előbbi az #1 szintnek felel meg, az utóbbi a #2 szintnek illetve a #3 szint nem megfelelő betűkre vonatkozó első felének. Az eredendően betűkre, betűpárookra vonatkozó szabályokat tartalmazó szabályrendszert kiegészítettük az országnévek fordításával, valamint a településnevek végén gyakran előforduló elemek (pl.: -falva, -háza) helyreállított alakjának listájával.

Minden adatmezőhöz tartozik egy elvárt értékeket (vezetékneveket, keresztneveket, országokat, településeket stb.) tartalmazó gazetteer-lista, egy opcionális gyakorisági („fontossági”) lista, valamint egy szabályrendszer a fenti két verzióban. Ezt egyben az adatmezőhöz tartozó *eszközkészlet*nek nevezzük.

6.3. Az algoritmus

A helyreállítást megvalósító algoritmus a következő lépésekből áll.

1. Előkészítjük az adatmezőhöz rendelt eszközkészletet.
2. Átírjuk az adatot a laza átíróval.
3. Az így kapott reguláris kifejezést illesztjük a listára.
4. Megtalálható a listán így? Ha igen, visszaadjuk az összes találatot. ✓
5. Ha nincs, akkor közelítő kereséssel keressük a szigorú átíratot a listán.
6. Megtalálható? Ha igen, visszaadjuk a legjobb találatot. ✓
7. Egyébként: visszaadjuk a szigorú átíratot. ✓

Az algoritmusnak három (pipával jelölt) kimeneti pontja van, ezek rendre megfelelnek az #1+#2 szintnek, a #3 szintnek illetve annak az esetnek, amikor semmilyen módon nem sikerült a megfelelő lista egy elemeként azonosítani a helyreállított alakot, így a pusztá szigorú átíratnál jobbat nem tudunk mondani. 'Имре' → Imre esetén már az #1 szint eredményt adna. Ha 'Андром' a kiinduló adat, akkor reguláris kifejezés segítségével találjuk meg a helyes András alakot (vö: 2. táblázat). 'Ференц' bemenetnél a regex nem segít, mert az orosz 'о'-nak a laza átíró nem felelteti meg az e-t. Itt a közelítő keresés találja meg a Ferenc alakot. Végül, ha a 'Момольсильтер' szóalak az adat, akkor egyik megközelítés sem jár sikerrel, így csupán a szigorú átíróval létrehozott Momolsilter alakot tudjuk visszaadni.

A módszert python nyelven implementáltuk, a 5. lépésben a közelítő keresést a `difflib` csomaggal valósítottuk meg. Az 4. lépésben a kapott találati halmaz sorrendezéséhez két szempontot veszünk figyelembe: egyrészt a találatoknak a szigorú átíratához való hasonlóságát, valamint lehetőség szerint a találatok általános gyakoriságát. Előbbit a `difflib` megfelelő függvényével számítjuk ki, utóbbi adat a vezetéknevek és a keresztnévek esetében állt a rendelkezésünkre egy első világháborús veszteséglista formájában. A találatokat a két adatból képzett közös pontszám sorrendjében, a pontszámmal együtt adjuk vissza. Az 'Андром' alakból laza átíróval képzett regex – $(A|\bar{A})(n|m)(d|gy|t)(r|l)(a|\bar{a}|o|e)(s|sch)$ – 192 különböző alakot fed le. Erre a lazaságra általában szükség is van, mert az egyes adatelemek nagyon sok változatban valóban előfordulnak.

6.4. Idegennyelvű szövegek

Az eredeti kiindulópontunkkal szemben, miszerint *magyar* nyelvű adatokat kell helyreállítanunk, kitűnt, hogy egyéb célnyelv (vö: 3. rész) is előfordul. Például a német. Egyrészt számos német nemzetiségű hadifogoly is volt, másrészt sokakat akkor fogtak el, mikor a front már Ausztria területén járt, így az elfogás helye osztrák település (6. táblázat).

A német adatelemek kezelésére szükséges volt létrehozni egy komplett *orosz-német* átíró szabályrendszert, amiben a németnek megfelelő szabályok kaptak helyet: 'ц' → c helyett 'ц' → z; 'в' → v helyett 'в' → w; 'аѣ' → a helyett 'аѣ' → ei; 'оѣ' → o helyett 'оѣ' → eu stb.

Гроц	→ Graz
Лииц	→ Linz
Фрайштадт	→ Freistadt
Дойчландберг	→ Deutschlandsberg
Штокаров	→ Stockerau
Цветел	→ Zwettl

6. táblázat. Osztrák településnevek az adatbázisban.

A rendszer működését alkalmassá tettük arra, hogy bizonyos feltétel teljesülése esetén alternatív eszközkészlet használatára lehessen váltani. Azon helyek esetén tehát, melyekben szerepel az 'Австрия' alak, átváltunk az orosz-német szabályok + osztrák településlista eszközkészletre.

7. Eredmények, megfontolások

7.1. Kiértékelés

Az előzőekben bemutatott eljárás kiértékelésére két mértéket használtunk. A *megalapozott helyreállítások aránya* (M) egy fedés jellegű mérték, azt mutatja meg, hogy az adatok hány százalékára tud a módszerünk a buta szigorú átírásnál jobb megoldást adni (ld. az algoritmus 4. és 6. pontja a 6.3. részben). A *helyes helyreállítások aránya* (H) egy pontosság jellegű mérték, azt mutatja meg, hogy az adatok hány százalékára van valóban helyes helyreállítás. A nagyon specifikus feladat miatt más módszerekkel való összevetésre nincs lehetőség. Az M és H értékeket a 7. táblázat mutatja be. Az adatok kezelhetőség szempontjából az M érték alapján kétféle oszlanak: a nevek és az országok (✓a táblázatban) jól kezelhetők ($M = 95\text{--}100\%$), a többi helyadat sokkal nehezebb ($M = 50\text{--}70\%$). Utóbbiban valószínűleg közrejátszanak a felbontás nehézségei (vö: 6.1. rész).

A H értékeket is figyelembe véve négy csoport látszik. A keresztnévek és az országnevek (2-4. és 8. sor) szinte mindegyikére van ajánlata az algoritmusnak és lényegében az összes ajánlat helyes is. A vezetéknéveknél (1. sor) a helyesség alacsonyabb. Közrejátszik, hogy keresztnévekhez képest sokkal ($\sim 50\times$) több vezetéknév van, jóval több nehezen megfejthető példa fordul elő: 'Хымаро' → Homoga?, 'Турул' → Turul?; valamint, hogy a *B. Kovács* típusú összetett vezetéknévek nincsenek jelenleg kezelve. A megyéknél (5. sor) kevesebb helyen tud tippet adni az algoritmus, olyankor viszont szinte mindig helyes a tipp. Itt jellemző hiba, hogy keverednek a különböző méretű közigazgatási egységek: *Derecske* vagy *Gyömrő* például megjelenik megyeként is az adatbázisban. A település és a járás (6-7. és 10-11. sor) a legnehezebb. A járás kisebb fontosságú és eléggé ritkán is fordul elő, a település viszont kiemelten fontos a hadifoglyok azonosításhoz. Itt sajnos az M és a H is alacsony, leginkább ezen szükséges javítani a jövőben. A nehezen megfejthető példák (pl.: Фоло, Улалануш) mellett gondot jelentenek itt a nem-osztrák külföldi települések, valamint a gyakran előforduló hosszú te-

adatmező	M	helyes / összes =		H	H/M
✓ 1. vezetéknev	95,8%	76	100	76%	79%
✓ 2. keresztnév	95,0%	92	100	92%	97%
✓ 3. apai keresztnév	95,2%	70	78	90%	95%
✓ 4. születés: ország	99,9%	45	45	100%	100%
5. születés: megye	70,7%	32	49	65%	92%
6. születés: járás	60,9%	7	15	47%	77%
7. születés: település	65,9%	31	61	51%	77%
✓ 8. fogságba esés: ország	99,9%	33	33	100%	100%
9. fogságba esés: megye	67,7%	11	12	92%	—
10. fogságba esés: járás	46,2%	1	4	25%	54%
11. fogságba esés: település	67,5%	29	56	52%	77%
összesen	85,5%	427	553	77%	90%

7. táblázat. Eredmény: az adatok 77%-ához tudunk helyes helyreállított alakot rendelni. A kiértékelésben a megalapozott helyreállítások aránya (M , „fedés”) és a helyes helyreállítások aránya (H , „pontosság”) szerepel. Előbbit a teljes, 682000 rekordot tartalmazó adatbázis alapján számoltuk, utóbbit az adatbázisból képzett 100 rekordból álló random mintán állapítottuk meg manuális kiértékeléssel. Az *összes* mező mutatja, hogy 100 sorból hányban volt jelen a szóban forgó adat. A H/M arány arról informál, hogy a megalapozott helyreállítások hány százaléka helyes valóban. A 9. sorban – vélhetően a H -hoz használt minta kis mérete miatt – nem értelmezhető érték adódik.

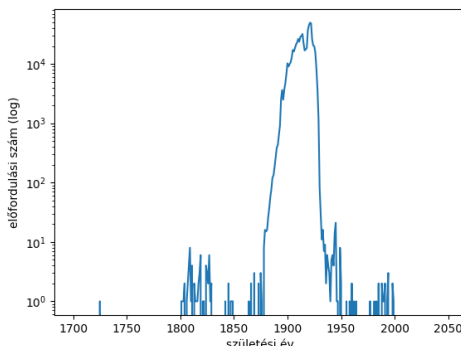
lepülésnevek, melyek gyakran számos hibát tartalmaznak (pl.: 'Яскорогенуї' → Jászkarajenő?, 'Пишпекляний' → Püspökladány?).

Látjuk, hogy az M és H értékek sok helyen összecsengenek: ahol tud valamit mondani az algoritmus, ott legtöbb esetben helyes is a javasolt helyreállítás. Az M -kiértékelés automatikus (össze kell számolni), a H -kiértékelés manuális munkát igényel. A H/M arány megmutatja, hogy az algoritmus által szolgáltatott megoldások mennyire jók. Ha ez magas, annak az az előnye, hogy megspórolhatjuk a munkaigényes H -kiértékelést, mert ekkor a H érték jól becsülhető az M értékkel. Ez az eset azokra az adatmezőkre jellemző, ahol lehetséges helyes adatok száma alacsony.

7.2. A szabadszöveges adatbázisokról

A kiinduló adatbázisunk egy *kézzel készült, korlátozatlan, szabadszöveges* adatbázis. Ez azt jelenti, hogy az adatbevitel nincs semmilyen értelemben korlátozva – például legördülő menüből való választás vagy típusellenőrzés révén –, azaz teljesen szabadon azt ír be az adatmezőkbe, amit csak akar. Az ilyen adatbázisok szükségszerűen következtelenek, mivel nincs olyan mechanizmus, ami biztosítaná az adatok egységességét: hogy ugyanazt mindig ugyanúgy jelöljük, az eltérő dolgokat pedig mindig eltérően.

Amellett, hogy az ilyen adatbázisokban egy adat több validnak mondható formában fordul elő, az ilyen adatbázisokba ellenőrzés híján számos hiba, elírás is belekerül. Azt látjuk, hogy ha nem legördülő menüből kell választani, akkor még a születési év adatot is el lehet rontani (2. ábra). A tanulság az, hogy az adatbázisok készítésekor szükséges az ellenőrzés, az egységesítő mechanizmus.



2. ábra: A adatbázisban szereplő *születési év* adatmező értékeinek eloszlása. A második világháborús hadifoglyok adatai között előfordul 1725-ös és 1999-es születési év is.

Visszont van olyan eset is, amikor valóban szabad kezet akarunk adni az adat-szolgáltatónak/adatrögzítőnek. Véleményünk szerint ilyen eset a közvélemény-kutatás. Azt gondoljuk, hogy ha egy közvéleménykutatási kérdés esetén – főként ha *miért*-es kérdésről van szó – a válaszadónak néhány előre megadott választási lehetőség közül kell választania, akkor a kutatás szükségszerűen veszít a hitelességéből ahhoz képest, ha a válaszokat szabadon fogalmazhatja meg, például mivel adott esetben véleményét jól visszaadó válasz egyszerűen nem szerepel a lehetőségek között.

A korlátozott módon készülő adatbázisokat persze sokkal könnyebb kiértékelni. Ugyanakkor a korlátozatlan adatbázisok adatainak értelmezése is megvalósítható: nyelvtechnológiai eszközökkel. Két esetben kaphat tehát szerepet a nyelvtechnológia: amikor nem történt előzetes adatellenőrzés/korlátozás (pl. a jelen tanulmányban tárgyalt hadifogoly-adatbázis) illetve amikor nem akarunk előzetes adatellenőrzést/korlátozást (pl. közvéleménykutatás).

A szabadszöveges adatbázisok értelmezési-feldolgozási munkálatait érdemes három szakaszra bontani: (1) adatvizsgálaton alapuló szakasz; (2) gyakorisági hibaelemzésen alapuló szakasz; (3) manuális szakasz. Az első szakaszban valamilyen automatikus rendszer áll elő, ami az adatok jelentős részét kezelni képes, a második szakaszban ezt finomítjuk a felfedett gyakori hibák javítása révén. Tudva azt, hogy ha a tökéleteshez közeli eredményt szeretnénk, akkor nem le-

het megspórolni a manuális szakaszt, a második szakaszban azokkal a hibákkal foglalkozunk, amelyek javítása a legnagyobb haszonnal jár.

A jelen tanulmányban feldolgozott adatbázis a többrétű torzulás miatt a szabadszöveges adatbázisoknak is a szélsőségesen következetlen és sokféle hibával teli fajtájához tartozik. Kezeléséhez a fent (6. rész) ismertetett szabályalapú megközelítést alkalmaztuk. Azért fogtunk hozzá így, mert egyrészt egy teljesen egyedi feladat konkrét problémáit kellett megoldani behatárolt méretű adathalmazon, valamint tanulóadat híján a gépi tanulási módszerek alkalmazására nem volt lehetőség. Ilyenkor ma is lehet létjogosultsága a szabályalapú módszereknek.

7.3. Példák

A 8. táblázatban egy engedéllyel közzétett valódi teljes példa látható.

vezetéknév	Галь	Gál
keresztnev	Тибор	Tibor
apai keresztnev	Эмил	Emil
születési év	1915	1915
születés helye	г. Сольнок ул. Санопи, 17	Szolnok település, Szanopi (?) utca 17.
fogságba esés helye	г. Цветел, Австрия	Ausztria, Zwetel település
nemzetiség	венгр	magyar
fogságba esés ideje	12.05.1945	12.05.1945
elbocsátás ideje	08.07.1947	08.07.1947
fogadó tábor	сдан лагерь № 36	36-os tábor
rendfokozat	лейтенант	hadnagy
elbocsátó tábor	лагерь № 313	313-as tábor

8. táblázat. A helyreállító rendszer által kezelt adatok vastagítva láthatók. Egy helyen nem tökéletes a megoldás: a *Zwetel* helyesen *Zwetl* lenne.

Annak illusztrálására, hogy valóban előfordulhattak félreolvasási hibák (vö: 4. oldal) a kartonok elektronikus rögzítésekor, bemutatunk egy eredeti kartont (3. ábra).

A végső manuális szakaszra maradó adatok helyreállításának nehézségét két példán mutatjuk be. Ilyenkor előfordul, hogy egy-egy adat megfejtése önmagában kutatómunkát illetve több kutató együttműködését igényli. A 4. táblázat 8. bejegyzéseként látható Блодентмигайн helyreállítva Búdszentmihály. Itt kezelni kell a szó végén lévő 'йн' variációt, a kieső *sz*-t, valamint meg kell fejteni, hogy hogyan változhatott az *ї* az orosz 'лю' betűkapcsolattá. Ez az eset mindkét torzulástípust példázza, ugyanis minden valószínűség szerint az adatrögzítéskor lett *ї*-ből 'ю' megfelelőbb orosz betű híján; majd a digitalizáláskor 'ю'-ból 'лю' félreolvasás révén. A 6.3. részben idézett Момольсильтер helyreállítva Mosonszentpéter. Itt az segített, hogy az adott napon Mosonszentpéteren esett fogságba

КАРТОЧКА ИНТЕРНИРОВАННОГО Форма № 2

Под *интерния* Батальон № *1837* Рота №

1. Фамилия *Богачев* 3. Отчество *Николаевич*

2. Имя *Николай* 5. Место рождения *Варзуга*

4. Год рождения *1921* 6. Последнее место жительства *Уезд Тарту, с. Сари*

7. Национальность *русские* 8. Вероисповедание *уезд Тарту, ч. Таба № 26*

9. Партийность *нет* 10. Подданство (гражданство) *никому*

11. Профессия и специальность *20. Тренинг*

12. Образование: а) Общее *8кл* б) Специальное *32* в) Военное

13. Дата интернирования « *24 июля* 1945 года

3. ábra: Egy eredeti karton. Az írás értelmezése nagy gyakorlatot igényel.

a hadifoglyok nagy része, és az adatbázisban szerepeltek az idézett orosz alakra sokban hasonlító, de könnyebben megfejthető verziók is.

Köszönjük Nyéki Bence, Orosz Ferenc, Beke Gábor és Szatucsek Zoltán közreműködését a munkálatokban, illetve hozzájárulásukat a fenti példák megoldásához.

8. Elérhetőség

A helyreállító rendszer részletes technikai információkkal, teljes szabályrendszerekkel, listákkal, futtatható programmal és minden egyébvel elérhető a <https://github.com/dlt-rilmta/hadifogoly-adatbazis> címen. Az eredeti adatokat ez a repozitórium nem tartalmazza, rendelkezésre áll viszont egy adatmezőnként külön-külön randomizált adatfájl (*data/pseudo_1000_42.csv*), amin az eljárás a `make transcribe FILE=pseudo_1000_42` paranccsal futtatható.

9. Továbbfejlesztési lehetőségek

Az ismertetett módszer nem oldja meg maradéktalanul a kitűzött feladatot. Konkrét feladat lévén az elvi cél a teljes, 100%-os megoldás, ehhez, ahogy említettük, mindenképpen szükséges egy manuális munkaszakasz is. Ugyanakkor a teljesítmény növelése érdekében számos továbbfejlesztési lehetőség adódik.

A településlisták jelenleg magyar és osztrák településeket tartalmaznak. Sokszor előfordulnak hiányzó települések. A „magyar verziókat” (Bécs, Bukarest stb.) a magyar (átírószabályokhoz tartozó) listához szükséges hozzátenni, az idegen nyelvűeket külön listákba szükséges gyűjteni, és a 6.1. rész elején említettek miatt nyelvenként kezelni. Ehhez minden nyelvhez új átíró szabályrendszer szükséges, valamint egy módszer az aktuális nyelv meghatározására.

Megfontolható, hogy hogyan lehetne automatizálni a fogságba esés idejére alapuló a 7.3. rész legvégén említett ötletet a településnevek jobb azonosítására.

Érdemes lehet újból felmérni a hagyományos vagy neurális gépi tanulás alkalmazhatóságát. Manuális úton természetesen lehet tanítóadatot gyártani, ezzel eddig nem foglalkoztunk. A kiinduló adat ismertett extrém tulajdonságai miatt azonban nem vagyunk meggyőződve arról, hogy a gépi megközelítés esetünkben jelentősen jobb eredményt adna.

10. Összefoglalás

Tanulmányunkban egy orosz-magyar átíró és helyreállító rendszert mutattunk be, melynek célja a magyar hadifoglyok adatbázisából kiindulva értelmezni és helyreállítani az eredeti magyar adatokat a cirill betűs forma alapján. A kidolgozott szabályalapú módszer az adatok 77%-ához tud helyes helyreállított alakot szolgáltatni. Ez a kiinduló adatbázis minőségét tekintve megfelelő eredmény. A teljesítmény növelésére elsősorban a helyadatok feldolgozásában nyílik lehetőség a jövőben.

Az adatbázis jól példázza a kézzel készült, korlátozatlan, szabadszöveges adatbázisok szükségyszerű következtelenségét; az automatikus feldolgozást esetünkben még az adatok kétszeres torzulása is nehezíti. Az ilyen adatbázisok feldolgozását éppen a nyelvtechnológiai eszközök használata tudja megfelelően támogatni.

Hivatkozások

- Bradley, J.: The Mari web project's orthography helper(s) (2020), <https://www.univie.ac.at/maridict/site-2014/transcription-general.php>
- Haader, L.: Az ómagyar kódexek hibatipológiájának kutatásáról. In: Korompay, K., Stemler, Á., Terbe, E., C. Vladár, Z. (szerk.) Forráskutatás, forráskiadás, tudománytörténet II., pp. 23–33. Magyar Nyelvtudományi Társaság (2014)
- Katona, P., Szikla, G.: A Magyar Nemzeti Levéltár Hajdú-Bihar Megyei Levéltár adatbázisai. Új Nézőpont 1, 61–81 (2014)
- Morse, S.P.: Searching the Gulag database in one step (2005), <https://stevemorse.org/russian/gulag.html>
- Prószéky, G., Naszódi, M., Kis, B.: Recognition assistance - treating errors in texts acquired from various recognition processes. In: COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes. pp. 1263–1267. Taipei, Tajvan (2002)
- Zoltán, A.: A cirillbetűs írású szláv nyelvek szavainak és neveinek magyar helyesírása. MTA I. Osztály Közleményei 32, 171–192 (1981)

emPhon: Morphologically sensitive open-source phonetic transcriber

Kulcsár Virág^{1*}, Lévai Dániel^{12*}

¹HAS-BME Lendület Language Acquisition Research Group

²Alfréd Rényi Institute of Mathematics
{kulvirag@gmail.com, levoid@renyi.hu}

Abstract. We propose a new emtsv module which can provide phonetic transcription based on the emMorph state-of-the-art morphological analyzer. In the first part of the paper, we present the motivation, the main problem and the method we are using to leverage Hungarian phonetic transcription with the use of morphological analyses. The second part is about evaluation both intrinsically and extrinsically – we evaluate our transcriber based on the IPA transcriptions of Wiktionary and as part of a speech synthesizer system. The code and models are fully open-source and are available under LGPL 3.0 license at <https://github.com/levoid/emPhon>.

Keywords: phonetics, IPA, emtsv, morphology, analysis, speech, synthesis

1 Introduction, motivation

Hungarian is a highly phonetic language in its orthography. Compared to English, where the lack of productive morphology and the highly irregular orthography makes it hard for any rule-based transcriber to work efficiently, Hungarian orthography is much more informative. For a great extent, simple words like ‘alma’ *apple* are easy to phonetically transcribe: we can look up the IPA for Hungarian and just substitute the letters one-by-one, giving /ɒlmo/ as pronunciation. In most cases, however, it is needed to take into account the Hungarian phonology - there are large set of rules which are governed by not only the interaction of different phones, but by the morphology as well.

The simplest problem that a phonetic transcriber has to overcome in Hungarian is the handling of digraphs (e.g. *cs*, *ny*) and their long counterparts. This is not a trivial task in itself, even though it looks like an innocent problem. One could argue that a modern subword tokenizer is easily able to distinguish between digraphs and and co-occurrences, but that is simply not true. Consider the following example, where the *sz* has been segmented into *-s_ z-*: *als_zo_tok* ‘sleep-2PL’, produced both by HuBERT’s tokenizer (Nemeskey, 2020) and by XLM-Roberta’s (Conneau et al., 2020) tokenizer. This example demonstrates

* equal contribution

well that rule-based methods can indeed be better in a (decreasingly small) number of NLP tasks than neural or unsupervised methods.

The more complex problems arise when the morphology creates letter clusters which are pronounced differently based on the underlying morphology. Let us take the sentence *Tűnj el, tűnj el te vagy az igazi bűnjel*¹ and inspect the *-nj-* segments in *tűn-j* ‘disappear-2SG.IMP’ and *bűn-jel* ‘crime-sign.NOM’. The *-nj* segment at the end of *tűn-j* is pronounced as a voiced palatal nasal /ɲ/ due to the palatal assimilation in Hungarian. However, this rule does not hold in case of an inner (or an outer) word boundary: in *bűn-jel*, the *-nj-* cannot assimilate into /ɲ/, thus pronounced as /ɲj/ – the phones must be pronounced separately due to the previously mentioned inner word boundary.

A grapheme-to-phoneme transcriber which is able to solve the previously mentioned problems would be quite beneficial in speech systems or in phonetic research for example, and there is a lack of publicly available, open-source tool of this kind.

Given the above reasons and the versatility of the emtsv framework, we propose a new emtsv (Indig et al., 2019) module which would be of great use to the Hungarian NLP community.

2 Related work

Hungarian phonology, and more specifically, the interaction of pronunciation and orthography is a well-research subject. There are earlier, rule-based systems, even for Hungarian (Novák and Siklósi, 2016)², and there are machine learning-based methods, nowadays mostly neural-based techniques (Yolchuyeva et al., 2019), we have chosen to implement a rule-based system based on the theory of Hungarian phonology presented in Siptár and Törkenczy (2000).

Implementing those rules is not straightforward, since there is a lot of variation based on speed, style, dialect, and on the speaker’s unique speech.

There is also a problem with the International Phonetic Alphabet for Hungarian: it is very inconsistent, as further explained in section 2.1.

Due to these obstacles, we have decided for a middle-ground approach, where the output produced by our transcriber represent the pronunciation generally well but may fail to cover edge-cases. In the final version of the emPhon, however, we are including these rules as user-configurable, as to customize and to adapt to the domain in which the tool is used.

2.1 IPA for Hungarian

There are multiple versions of Hungarian IPA in circulation with various levels of accuracy. In this subsection, we go through the most distinctive differences of the transcriptions.

¹ <https://zeneszoveg.hu/dalszoveg/31311/kontroll-csoport/keresnek-zeneszoveg.html>

² Sadly, neither the model nor the data is available from this paper, so we are unable to compare our performance to theirs.

In the vowel system, there is one difference between various papers and IPA that is worth noting. Some consider the grapheme *a* to be a low back rounded vowel /ɒ/ (e.g. Szende (1994)), while others argue that it is a low-mid back rounded vowel /ɔ/ (e.g. Siptár and Törkenczy (2000); Gósy (2004)). Considering that we measure on Wiktionary data, we chose it to be low back, but it can easily be substituted in the final output with any text processing tool.

The system of consonants is more complicated. There is a multitude of problems which could be addressed here and above all of that, a lot of rules are dependent on the style, speed, and speaker. We will focus on the most frequent and practical difficulties that we have encountered.

Concerning the stop-fricative clusters, we do not distinguish these from affricates, giving us *cím* /t͡sim/ ‘address, title’ and *rendszer* /rent͡ser/ ‘system’, with *dsz* and *c* having identical representations. Even Wiktionary does not agree with itself; one³ Wiktionary page distinguishes between these, while another⁴ page does not.

Another problem in Hungarian IPA, the lack of marker for nasalization, even though it clearly happens, as in *tonhal* ‘tuna’ is represented as /tonhal/, instead of /tõ:hal/.

The ambiguity of Hungarian IPA will further be addressed in section 4.

3 Transcriber

Our code can be divided into two distinct processes. The first takes a morphologically analyzed text and creates segmented text with special delimiters. The second is where the actual transcribing is happening. In this step we create an inner representation for ease of regex matching and as the last step, we convert it into Hungarian IPA.

3.1 Morphological segmentation

Our input is morphologically analyzed text. For this we used the morphological analyzer of the emtsv framework (emMorph, Novák et al. (2016)). The output of emMorph is every possible analysis for the given word. Unfortunately there is no disambiguation between lemma analyses, so our workaround was to use the *finest* splitting. We defined the *finest* splitting as the one containing the most morphs (for example *ár_ad_ás* is finer than *árad_ás*, even when the former is highly unlikely) and that has proven to be acceptable baseline.

Segmentation was needed to avoid the pitfalls presented by words such as *kilenc-száz* ‘nine-hundred’ or *pác-só* ‘marinade-salt’. Here the letter ‘csz’ could be divided as *c-sz* or *cs-z* and while native speakers can easily decide which one is correct, the computer is lost without additional help.

³ https://en.wiktionary.org/wiki/Appendix:Hungarian_pronunciation

⁴ <https://en.wikipedia.org/wiki/Help:IPA/Hungarian>

That is why we enlisted four different separators. The characters used are: |, #, §, ~, and they all signify different types of transitions that can be seen in table 1.

Sep.	Transition	Example
#	between roots in compound words	rend#szer
~	between root and first suffix	emészt~és
§	between prefix and root	ki§mond
	between affixes	anyag~ok at

Table 1. Separators and their transitions

By default, having four different separators is superfluous, but there might be edge-cases where the extra information provided by these separators can prove to be useful.

3.2 Transcription rules

After morphologically segmenting the text we can start transcribing. We based our pronunciation rules on Siptár and Törkenczy (2000), and implemented most of them using regex. This keeps the code pure python and provides a smoother installation experience while keeping the coding rather straightforward and running time low.

To handle digraphs and long phonemes, we created a unicode inner representation where every character defines exactly one sound. We decided to use the uppercase letters to stand for geminates in our inner representation. The double letter function has to be the first one to run in order to properly preprocess text. The long letter function runs as the finishing step of every other function so if another rule creates double consonants, they are handled. Below is the overview of different rules. To see the exact rules, the similarly named functions in the code should be referenced. The tool is implemented as a Python class with methods as rules, thus the order and the need for multiple runs can be easily achieved.

- **Digraphs:** The digraphs are substituted at the beginning of the process.
- **Degemination:** There are 4 different types of geminates based on the order of the geminate and the consonant and on whether the geminate is segmented or not.
- **Hiatus filling:** Hiatuses are filled in with /j/ in Hungarian.
- **Nasal place assimilation:** The nasal consonants assimilate based on the location of the consonant after it.
- **Elision of n:** The /n/ fully assimilates with the non-nasal sonorants after.
- **Elision of l:** The /l/ fully assimilates if there is /r/ after.
- **Palatal assimilation:** There are some consonants which are sensitive to palatalization and if these are adjacent to a palatal, they interact.

- **Sibilant rules:** There are two larger sets of rules here. One treats the sibilant fricative clusters, the other treats stop + sibilant clusters.
- **/h/-alternations:** The alternation of /h/ is a very complex phenomenon. There is a high degree of variance from speaker-to-speaker, thus we aimed to maximize the coverage on our evaluation dataset using a very limited number of rules.
- **Voice assimilation:** In Hungarian, we have regressive voicing assimilation, meaning that the voice of an obstruent cluster is governed by the rightmost consonant.

The last step is to convert the inner representation into standardized IPA.

4 Evaluation

We evaluate our phonetic transcriber two ways: first, we compare it with gold-standard phonetic transcriptions downloaded from the Hungarian Wiktionary⁵. Second, we use it to leverage a text-to-speech system on a single-speaker dataset.

4.1 Wiktionary

Wiktionary is a sister-project of Wikipedia, it is a multilingual, free content dictionary, it is run by the Wikimedia Foundation and is written by volunteers. Extracting the phonetic data was not straightforward: apparently the phonetic transcriptions are not saved in the periodic Wiktionary dumps⁶, thus we had downloaded 65586 pages one-by-one and extracted the pronunciations using regular expressions.

The gold-standardness of this Wiktionary data is highly arguable based on our observations. There are systematic errors which indicate that there is some kind of automatic transcriber writing these. Below, we give an overview of the common errors.

One of the common errors is the creation of digraphs through morphological segments.

- vízsugár: /vi:ʒuga:r/
- gázság: /gɒʒa:g/

Using hiatus filling unnecessarily is also a common error.

- látóideg: /la:to:jideg/
- adójóváírás: /ɒdo:jor:va:jira:f/

The pronunciation of numbers is also questionable in places.

- harminchét: /hormintʃet/

⁵ <https://hu.wiktionary.org/>

⁶ <https://dumps.wikimedia.org/huwiktionary/>

- ezerkilencszázharminchat: /ezerkilen^gza:shorminfot/

There is a general problem with the inclusion of foreign proper nouns.

- Auschwitz: /oufxvidz/
- Schmah!l: /fxmaxl/

We have also encountered errors caused by some words redirecting to other words’ page, and therefore having the wrong pronunciation associated with it. This mostly happened with slang words of crude meaning. In order to avoid this we decided to dismiss the words which had a pronunciation considerably longer or shorter than the original word length. In the end we were left with 64666 data points.

Morph	h-deletion	WER
Yes	No	4.10 %
Yes	Yes	3.25 %
No	No	3.71 %
No	Yes	2.86 %

Table 2. Word Error Rates

It will be quite hard to achieve lower WER on Wiktionary due to the problems above. As there are systematic errors in the last few percents, one has to decide the target of optimization: the Wiktionary data or the actual pronunciation. Some of the remaining problems can be rectified by creating a pronunciation lexicon which details how words should be written to reflect Hungarian pronunciation.

We also measured the WER by eliminating the postlexical /h/ variants and rules both in the dataset and in the transcriber. This is justified by the complexity and marginality of the Hungarian /h/.

We also measured the effect of excluding the morphological information since the Wiktionary data does not appear to be sensitive to it. This resulted in a lower WER. The improvement can be mostly explained by the appearance of the aforementioned common errors, such as the incorrect digraphs and hiatus filling.

4.2 Speech synthesis

We used the Tacotron 2 system (Wang et al., 2017) to extrinsically evaluate our model. The Tacotron 2 model is a sequence-to-sequence model trained to map a sequence of letters to a sequence of features that encode the audio. These features are audio spectrograms computed on short segments of audio thus they capture not only the pronunciation, but also the volume, speed and intonation. This representation is then converted to waveform using a WaveNet-like architecture (van den Oord et al., 2016).

The lack of freely-accessible Hungarian speech data has limited our options, thus we have settled with the Hungarian part of the CSS10 dataset⁷ created by Park and Mulc (2019). It is essentially the annotated and segmented audiobook of *Egri Csillagok* ‘Eclipse of the Crescent Moon’⁸ read by a single speaker.

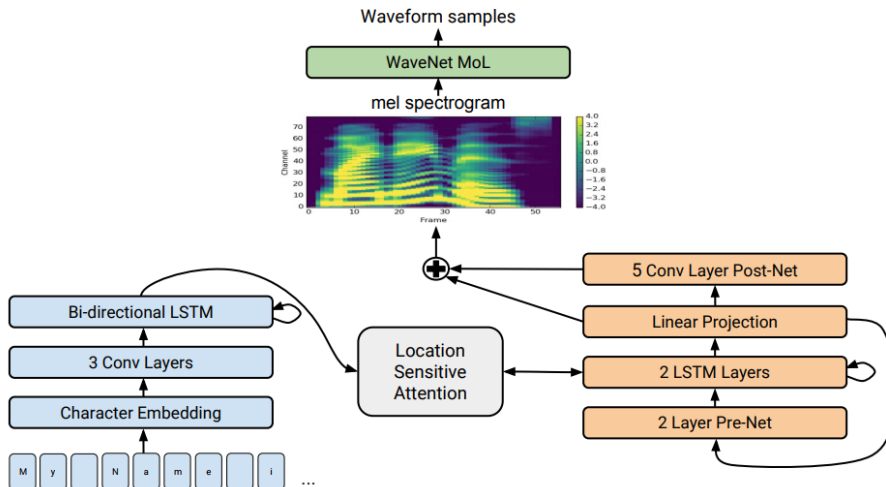


Fig. 1: Architecture of Tacotron

In our test setup, we used the Tacotron implementation of Nekvinda and Dušek (2020)⁹. We trained the Tacotrons for 300 epochs with the default hyperparameters presented in the Github repository, and we have used their pre-trained model from the Github repository of their WaveRNN¹⁰ implementation. We compare the models that were trained on the original text to the model trained on the phonetically transcribed text. The training of both models took around 1 day on a single 1080Ti GPU.

Initially, we expected the two models to be almost indistinguishable, since the labels (audio recordings) are the same, and the difference between the training datasets are not that significant. However, our impressions of the difference is that it is perceptible, even by the naïve ears, but this distinction is not easy to quantify.

The baseline model can already produce clean, natural, comprehensible speech, and our goal was to achieve similar performance to prove that our transcriber

⁷ <https://www.kaggle.com/bryanpark/hungarian-single-speaker-speech-dataset>

⁸ https://en.wikipedia.org/wiki/Eclipse_of_the_Crescent_Moon

⁹ https://github.com/Tomiinek/Multilingual_Text_to_Speech

¹⁰ https://github.com/Tomiinek/Multilingual_Text_to_Speech/releases/download/v1.0/wavernn_weight.pyt

does not lose information. From this point of view, we succeeded. Our model, which trained on the phonetic transcription, is by no means inferior to the baseline model. There is a subtle difference in intonation and surprisingly, in tone, but one of the starker differences was audible in case of /h/-s and affricates. The /h/-s are more pronounced and somewhat more natural in our model, which is expected since they are generally rare in occurrence but are influenced by many rules. On the other hand, our model makes less of a distinction between affricates and fricatives. In many cases, /tʃ/ and /s/ sound too similar. This can happen due to our tool merging too many *-tsz-* and *-dsz-* segments, when in reality, these are frequently pronounced differently from *-c-*.

5 Conclusion and further research

In this paper, we have presented an automatic phonetic transcriber tool which is publicly available, fast, and fits well the existing emtsv framework, thus it is easy to use. The small amount of publicly available data limited our ability to fully evaluate the model, but as we have shown in Section 4, the transcriber performs well.

There are multiple directions for further research. On one side, the transcriber’s accuracy can easily be improved with the usage of a curated pronunciation lexicon. As another direction, the speech synthesis models have shown that the transcriber’s performance can be ‘heard’, giving us an interesting insight into what happens when we force the system to learn based on the phonetic forms. We are certain that there are more interesting phenomena happening here which could be discovered by experienced phoneticians.

6 Acknowledgements

This work was supported by the Lendület Research Grant of the Hungarian Academy of Sciences (Lendület 96233 ‘Profiling learning mechanisms and learners: individual differences from impairments to excellence in statistical learning and in language acquisition’, PI: Ágnes Lukács). The work of Lévai is partially supported by NRDIForefront Research Excellence Program KKP_20 Nr. 133921, Hungarian National Excellence Grant 2018-1.2.1-NKP-00008, and by the Ministry of Innovation and the National Research, Development and Innovation Office within the framework of the Artificial Intelligence National Laboratory Programme.

Bibliography

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale (2020)

- Gósy, M.: Fonetika, a beszéd tudománya. Osiris Kiadó, Budapest, Magyarország (2004)
- Indig, B., Sass, B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: **emtsv** — egy formátum mind felett. In: Berend, G., Gosztolya, G., Vincze, V. (eds.) XV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2019). pp. 235–247. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2019)
- Nekvinda, T., Dušek, O.: One model, many languages: Meta-learning for multi-lingual text-to-speech (2020)
- Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)
- Novák, A., Siklósi, B.: Grapheme-to-phoneme transcription in hungarian. *Int. J. Comput. Linguistics Appl.* 7, 161–173 (2016)
- Novák, A., Siklósi, B., Oravecz, C.: A new integrated open-source morphological analyzer for Hungarian. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France (may 2016)
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. *CoRR abs/1609.03499* (2016), <http://arxiv.org/abs/1609.03499>
- Park, K., Mulc, T.: CSS10: A collection of single speaker speech datasets for 10 languages. *CoRR abs/1903.11269* (2019), <http://arxiv.org/abs/1903.11269>
- Siptár, P., Törkenczy, M.: The phonology of Hungarian. Oxford: Oxford University Press. *Phonology* 18(2) (2000)
- Szende, T.: Illustrations of the IPA: Hungarian. *Journal of International Phonetic Association* pp. 91–94 (1994)
- Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.V., Agiomyrgiannakis, Y., Clark, R., Saurous, R.A.: Tacotron: A fully end-to-end text-to-speech synthesis model. *CoRR abs/1703.10135* (2017), <http://arxiv.org/abs/1703.10135>
- Yolchuyeva, S., Németh, G., Gyires-Tóth, B.: Transformer based grapheme-to-phoneme conversion. *Interspeech 2019* (Sep 2019), <http://dx.doi.org/10.21437/Interspeech.2019-1954>

Automatic punctuation restoration with BERT models

Attila Nagy¹, Bence Bial¹, Judit Ács¹

¹Department of Automation and Applied Informatics
Budapest University of Technology and Economics

Abstract. We present an approach for automatic punctuation restoration with BERT models for English and Hungarian. For English, we conduct our experiments on Ted Talks, a commonly used benchmark for punctuation restoration, while for Hungarian we evaluate our models on the Szeged Treebank dataset. Our best models achieve a macro-averaged F_1 -score of 79.8 in English and 82.2 in Hungarian. Our code is publicly available¹.

1 Introduction

Automatic Speech Recognition (ASR) systems typically output unsegmented transcripts without punctuation. Restoring punctuations is an important step in processing transcribed speech. Tündik et al. (2018) showed that the absence of punctuations in transcripts affect readability as much as a significant word error rate. Downstream tasks such as neural machine translation (Vandeghinste et al., 2018), sentiment analysis (Cureg et al., 2019) and information extraction (Makhoul et al., 2005) also benefit from having clausal boundaries. In this paper we present models for automatic punctuation restoration for English and Hungarian. Our work is based on a state-of-the-art model proposed by (Courtland et al., 2020), which uses pretrained contextualized language models (Devlin et al., 2018).

Our contributions are twofold. First, we present the implementation of an automatic punctuation model based on a state-of-the-art model (Courtland et al., 2020) and evaluate it on an English benchmark dataset. Second, using the same architecture we propose an automatic punctuator for Hungarian trained on the Szeged Treebank (Csendes et al., 2005). To the best of our knowledge our work is the first punctuation restoration attempt that uses BERT on Hungarian data.

2 Related Work

Systems that are most efficient at restoring punctuations usually exploit both prosodic and lexical features with hybrid models (Szaszák and Tündik, 2019; Garg et al., 2018; Żelasko et al., 2018). Up until the appearance of BERT-like

¹ <https://github.com/attilanagy234/neural-punctuator>

models, lexical features were primarily processed by recurrent neural networks (Vandeghinste et al., 2018; Tündik et al., 2017; Kim, 2019; Tilk and Alu    , 2016; Salloum et al., 2017), while more recent approaches use the transformer (Vaswani et al., 2017) architecture (Chen et al., 2020; Nguyen et al., 2019; Cai and Wang, 2019). The current state-of-the-art method by Courtland et al. (2020) is a pretrained BERT, which aggregates multiple predictions for the same token, resulting in higher accuracy and significant parallelism.

3 Methodology

We train models for Hungarian and English. For English we rely on the widely used IWSLT 2012 Ted Talks dataset (Federico et al., 2012) benchmark dataset. Due to the lack of such datasets for Hungarian, we generate it from the Szeged Treebank (Cs       et al., 2005). We preprocess the Szeged Treebank such that it structures similarly to the output of an ASR system. Then with the presented methods we attempt to reconstruct the original and punctuated gold standard corpus.

3.1 Problem formulation

We formulate the problem of punctuation restoration as a sequence labeling task with four target classes: *EMPTY*, *COMMA*, *PERIOD*, and *QUESTION*. We do not include other punctuation marks as their frequency is very low in both datasets. For this reason, we apply a conversion in cases where it is semantically reasonable: we convert exclamation marks and semicolons to periods and colons and quotation marks to commas. We remove double and intra-word hyphens, however, if they are encapsulated between white spaces, we convert them to commas. Other punctuation marks are disregarded during our experiments. As tokenizers occasionally split words to multiple tokens, we apply masking on tokens, which do not mark a word ending. These preprocessing steps and the corresponding output labels are shown in Table 1.

3.2 Datasets

IWSLT 2012 Ted Talks dataset We use the IWSLT 2012 Ted Talks dataset (Federico et al., 2012) for English. IWSLT is a common benchmark for automatic punctuation. It contains 1066 unique transcripts of Ted talks with a total number of 2.46M words in the corpus. We lowercase the data and we convert consecutive spaces into single spaces. We also remove spaces before commas. We use the original train, validation and test sets from the IWSLT 2012 competition. The overall data distributions of the IWSLT Ted Talk dataset is summarized in Table 2.

Original	Tyranosaurus asked: kill me?					
Preprocessed	tyranosaurus asked, kill me?					
Tokenized	ty	##	rano	##	saurus	asked kill me
Output	-	-	EMP	COM	EMP	Q
Original	Not enough, – said the co-pilot – ...					
Preprocessed	not enough, said the co pilot,					
Tokenized	not	enough	said	the	co	pilot
Output	EMP	COM	EMP	EMP	EMP	COM

Table 1: An example input sentence and the following processing steps in our setup.

	Train	Validation	Test
PERIOD	139,619	909	1,100
COMMA	188,165	1,225	1,120
QUESTION	10,215	71	46
EMPTY	2,001,462	15,141	16,208

Table 2: Label distributions of the IWSLT Ted talk dataset.

Szeged Treebank We use the Szeged Treebank dataset (Csendes et al., 2005) for Hungarian. This dataset is the largest gold standard treebank in Hungarian. It covers a wide variety of domains such as fiction, news articles, and legal text. As these subcorpora have very different distributions in terms of punctuations, we merge them and shuffle the sentences. We then split the dataset into train, validation and test sets. This introduces a bias in the prediction of periods as it is easier for the model to correctly predict sentence boundaries by recognizing context change between adjacent sentences but it also provides a more-balanced distribution of punctuation classes across the train, validation and test sets. The label distribution is listed in Table 3.

3.3 Architecture

Our model is illustrated in Figure 3. We base our model on pretrained BERT models. BERT is a contextual language model with multiple transformer layers and hundreds of millions trainable parameters trained on a massive English corpora with the masked language modeling objective. Several variations of pre-trained weights were released. We use BERT-base cased and uncased for English as well as Albert (Lan et al., 2019), a smaller version of BERT. BERT also has

	Train	Validation	Test
PERIOD	81,168	9,218	3,370
COMMA	120,027	13,781	4,885
QUESTION	1,808	198	75
EMPTY	885,451	101,637	36,095

Table 3: Overall data distributions of the Szeged Treebank dataset.

a multilingual version, *mBERT* that supports Hungarian along with 100 other languages. We use mBERT and the recently released Hungarian-only BERT, *huBERT* (Nemeskey, 2020) for Hungarian. These models all apply wordpiece tokenization with their own predefined WordPiece vocabulary. They then generate continuous representations for every wordpiece. Our model adds a two-layer multilayer perceptron on top of these representation with 1568 hidden dimension, ReLU activation and an output layer, and finally a softmax layer that produces a distribution over the labels. We also apply dropout with a probability of 0.2 before and after the first linear layer. Similarly to Courtland et al. (2020), we apply a sliding window over the input data, generate multiple predictions for each token and then aggregate the probabilities for each position by taking the label-wise mean and thus output the most probable label. The process is illustrated in Figure 1 and 2.

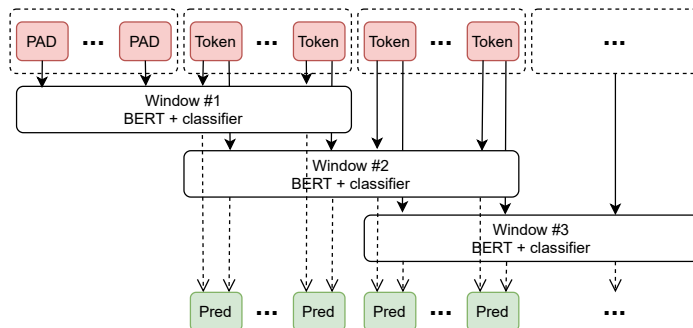


Fig. 1: The process of generating multiple predictions for a token. Although BERT always receives sequences of 512, we sample consecutive sequences from the corpora such that they overlap, thus resulting in multiple predictions for the same token. The extent of the overlap and therefore the number of predictions for a token depend on the offset between the windows. Note that padding is necessary in the beginning to ensure that all tokens have the same amount of predictions.

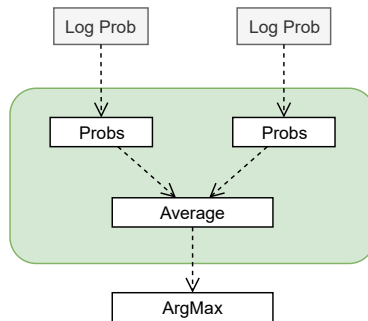


Fig. 2: The final prediction is computed by first aggregating all punctuation probability distributions for each token by taking their class-wise averages and then selecting the highest probability.

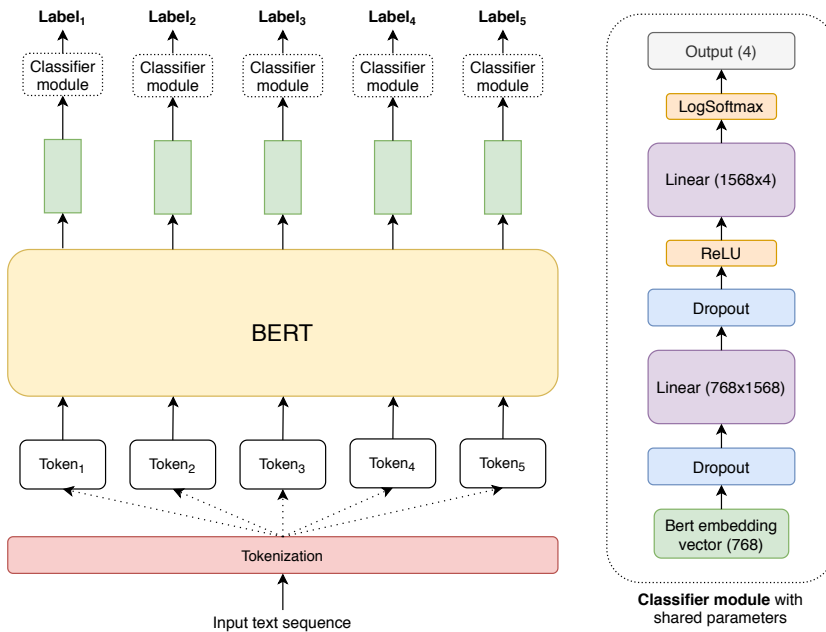


Fig. 3: The complete architecture used for punctuation restoration.

3.4 Training

We train all models with identical hyperparameters. We perform gradient descent using the AdamW optimizer (Loshchilov and Hutter, 2017) with the learning rate set to $3 * 10^{-5}$ for BERT and 10^{-4} for the classifier on top. We apply gradient clipping of 1.5 and a learning rate warm up of 300 steps using a linear scheduler. We select negative log likelihood as the loss function. The tokenizer modules often split single words to multiple subwords. For this task we only need to predict punctuations after words (between white spaces). We mask the loss function for every other subword. It is a common practice to intermittently freeze and unfreeze the weights of the transformer model, while training the fine-tuning linear layers situated at the top of the whole architecture. We found that it is best to have the transformer model unfrozen from the very first epoch and therefore update its parameters along with the linear layers. We trained the models for 12 epochs with a batch size of 4 and applied early stopping based on the validation set. We used the validation set to tune the sliding window step size, that is responsible for getting multiple predictions for a single token. All experiments were performed using a single Nvidia GTX 1070 GPU with one epoch taking 10 minutes. Our longest training lasted for 2 hours.

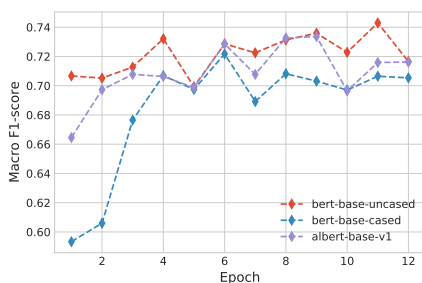
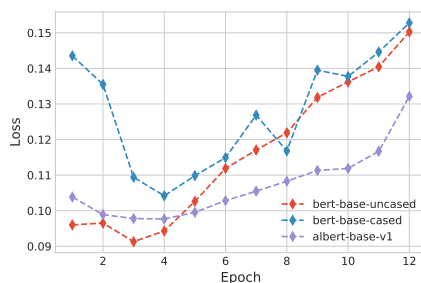
4 Results

All models are evaluated using macro F_1 -score (F) over the 4 classes. Similarly to Courtland et al., our work is focused on the performance of punctuation marks and as *EMPTY* labels constitute 85% of all labels, we report the overall F_1 -score without *EMPTY*. We evaluated both cased and uncased variations of BERT and generally we have found that the uncased model is better than its cased variant for this task. This was an expected conclusion, as we lowercased the entire corpus with the purpose of eliminating bias around the prediction of periods. For all setups, we selected the best performing models on the validation set by loss and by macro F_1 -score and evaluated them independently on the test set. On the Ted Talks dataset, our best performing model was an uncased variation of BERT that achieved on par performance with the current state-of-the-art model (Courtland et al., 2020), having a slightly worse macro F_1 -score of 79.8 (0.8 absolute and 0.9975% relative difference) with 10 epochs of training and 64 predictions/token. All results on the Ted Talks dataset are summarized in Table 4.

On the Szeged Treebank dataset, we evaluate the multilingual variants of BERT and the recently released Hubert model. We find that Hubert performs significantly better (82.2 macro F_1 -score) than the best multilingual model with an absolute and relative difference of 12.2 and 14.84% respectively on macro F_1 -score. We trained the best Hubert model for 3 epochs and used 8 predictions/token. All results on the Szeged Treebank dataset are summarized in Table 5.

We also examined the effect of using multiple predictions for a token. The changes we see in macro F_1 -score on the validation set with regard to the number

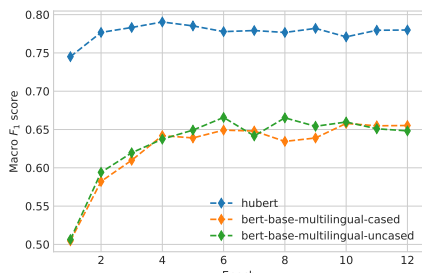
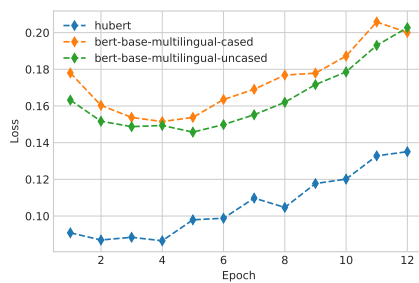
Models	Comma			Period			Question			Overall		
	P	R	F	P	R	F	P	R	F	P	R	F
BERT-base (Courtland et al., 2020)	72.8	70.8	71.8	81.9	86.6	84.2	80.8	91.3	85.7	78.5	82.9	80.6
Albert-base (Courtland et al., 2020)	69.4	69.3	69.4	80.9	84.5	82.7	76.7	71.7	74.2	75.7	75.2	75.4
BERT-base-uncased (by loss)	59.0	80.2	68	83.0	83.6	83.3	87.8	83.7	85.7	76.6	82.5	79.0
BERT-base-uncased (by F_1 -score)	58.4	80.7	67.8	84.2	83.8	84.0	84.8	90.7	87.6	75.8	85.1	79.8
BERT-base-cased (by loss)	57.3	73.9	64.5	75.9	87.9	81.4	77.1	84.1	80.4	70.1	81.9	75.5
BERT-base-cased (by F_1 -score)	59.1	78.5	67.5	79.6	81.6	80.6	76.9	88.9	82.5	71.9	83.0	76.8
Albert-base (by loss)	55.3	74.8	63.6	76.8	87.9	82.0	70.6	83.7	76.6	67.6	82.1	74.1
Albert-base (by F_1 -score)	56.5	80.3	66.3	80.7	80.8	80.8	80.4	84.1	82.2	72.5	81.7	76.4

Table 4: Precision, recall and F_1 -score values on the Ted Talk dataset.(a) Macro F_1 -score

(b) Loss

Fig. 4: Metrics on the validation set over epochs during training on the IWSLT Ted Talk dataset

Models	Comma			Period			Question			Overall		
	P	R	F	P	R	F	P	R	F	P	R	F
BERT-base-multilang-uncased (by loss)	82.3	79.3	80.8	79.6	88.3	83.8	43.2	21.3	28.6	68.4	63.0	64.4
BERT-base-multilang-uncased (by F_1 -score)	82.9	79.4	81.1	80.1	88.4	84.0	51.4	24.0	32.7	71.5	63.9	66.0
BERT-base-multilang-cased (by loss)	81.3	79.3	80.3	82.4	83.2	82.8	51.6	21.3	30.2	71.8	61.3	64.4
BERT-base-multilang-cased (by F_1 -score)	83.6	78.8	81.1	81.7	85.5	83.6	61.4	36.0	45.4	75.6	66.8	70.0
Hubert (by loss and F_1 -score)	84.4	87.3	85.8	89.0	93.1	91.0	73.5	66.7	69.9	82.3	82.4	82.2

Table 5: Precision, recall and F_1 -score values on the Szeged Treebank dataset.(a) Macro F_1 -score

(b) Loss

Fig. 5: Metrics on the validation set over epochs during training on the Szeged Treebank dataset.

of predictions per token are shown in Figure 6. The best models were evaluated on the test set and we found that having multiple predictions per token increased the F_1 -score by 5% in English and 2.4% in Hungarian.

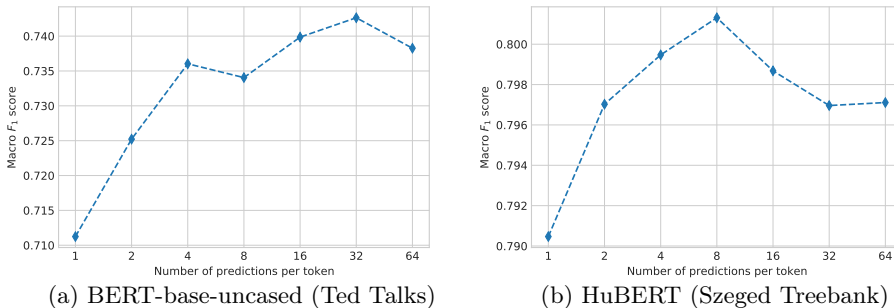


Fig. 6: Effect of the number of predictions per token on the overall F_1 -score, computed on the validation dataset.

5 Conclusion

We presented an automatic punctuation restoration model based on BERT for English and Hungarian. For English we reimplemented a state-of-the-art model and evaluated it on the IWSLT Ted Talks dataset. Our best model achieved comparable results with current state-of-the-art on the benchmark dataset. For Hungarian we generated training data by converting the Szeged Treebank into an ASR-like format and presented BERT-like models that solve the task of punctuation restoration efficiently, with our best model Hubert achieving a macro F_1 -score of 82.2.

Bibliography

- Cai, Y., Wang, D.: Question mark prediction by bert. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 363–367. IEEE (2019)
- Chen, Q., Chen, M., Li, B., Wang, W.: Controllable time-delay transformer for real-time punctuation prediction and disfluency detection. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8069–8073. IEEE (2020)
- Courtland, M., Faulkner, A., McElvain, G.: Efficient automatic punctuation restoration using bidirectional transformers with robust inference. In: Proceedings of the 17th International Conference on Spoken Language Translation. pp. 272–279 (2020)

- Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A.: The szeged treebank. In: International Conference on Text, Speech and Dialogue. pp. 123–131. Springer (2005)
- Cureg, M.Q., De La Cruz, J.A.D., Solomon, J.C.A., Saharkhiz, A.T., Balan, A.K.D., Samonte, M.J.C.: Sentiment analysis on tweets with punctuations, emoticons, and negations. In: Proceedings of the 2019 2nd International Conference on Information Science and Systems. pp. 266–270 (2019)
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
- Federico, M., Cettolo, M., Bentivogli, L., Michael, P., Sebastian, S.: Overview of the iwslt 2012 evaluation campaign. In: IWSLT-International Workshop on Spoken Language Translation. pp. 12–33 (2012)
- Garg, B., et al.: Analysis of punctuation prediction models for automated transcript generation in mooc videos. In: 2018 IEEE 6th International Conference on MOOCs, Innovation and Technology in Education (MITE). pp. 19–26. IEEE (2018)
- Kim, S.: Deep recurrent neural networks with layer-wise multi-head attentions for punctuation restoration. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7280–7284. IEEE (2019)
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
- Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
- Makhoul, J., Baron, A., Bulyko, I., Nguyen, L., Ramshaw, L., Stallard, D., Schwartz, R., Xiang, B.: The effects of speech recognition and punctuation on information extraction performance. In: Ninth European Conference on Speech Communication and Technology (2005)
- Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)
- Nguyen, B., Nguyen, V.B.H., Nguyen, H., Phuong, P.N., Nguyen, T.L., Do, Q.T., Mai, L.C.: Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. In: 2019 22nd Conference of the Oriental COCOSA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSA). pp. 1–5. IEEE (2019)
- Salloum, W., Finley, G., Edwards, E., Miller, M., Suendermann-Oeft, D.: Deep learning for punctuation restoration in medical reports. In: BioNLP 2017. pp. 159–164 (2017)
- Szaszák, G., Tündik, M.Á.: Leveraging a character, word and prosody triplet for an asr error robust and agglutination friendly punctuation approach. In: INTERSPEECH. pp. 2988–2992 (2019)
- Tilk, O., Aluăe, T.: Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In: Interspeech. pp. 3047–3051 (2016)

- Tündik, M.A., Szaszák, G., Gosztolya, G., Beke, A.: User-centric evaluation of automatic punctuation in asr closed captioning (2018)
- Tündik, M.Á., Tarján, B., Szaszák, G.: A bilingual comparison of maxent-and rnn-based punctuation restoration in speech transcripts. In: 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). pp. 000121–000126. IEEE (2017)
- Vandeghinste, V., Verwimp, L., Pelemans, J., Wambacq, P.: A comparison of different punctuation prediction approaches in a translation context. Proceedings EAMT (2018)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
- Želasko, P., Szymański, P., Mizgajski, J., Szymczak, A., Carmiel, Y., Dehak, N.: Punctuation prediction model for conversational speech. arXiv preprint arXiv:1807.00543 (2018)

SZEMANTIKA

Mitigating the Knowledge Acquisition Bottleneck for Hungarian Word Sense Disambiguation using Multilingual Transformers

Gábor Berend^{1,2}

¹University of Szeged, Institute of Informatics

²MTA-SZTE, Research Group on Artificial Intelligence

berendg@inf.u-szeged.hu

Abstract. A major hurdle in training all-words word sense disambiguation (WSD) systems for new domains and/or languages is the limited availability of sense annotated training corpora and that their construction is an extremely costly and labor-intensive process. In this paper, we investigate the utilization of multilingual transformer-based language models for performing cross-lingual WSD in the zero-shot setting. Our empirical results suggest that by relying on the intriguing multilingual abilities of pre-trained language models, we can infer reliable sense labels to Hungarian textual utterances in the all-word WSD setting by purely relying on sense-annotated training data in English.

Keywords: zero-shot word sense disambiguation; contextualized word representations; knowledge acquisition bottleneck

1 Introduction

A key difficulty in natural language understanding is definitely the highly ambiguous nature of natural language utterances. This property has made word sense disambiguation (WSD) a long-standing and central task within the NLP community (Lesk, 1986; Gale et al., 1992; Navigli, 2009) with ample application possibilities, e.g. in information retrieval (Zhong and Ng, 2012).

Most successful WSD systems are built in a monolingual and supervised manner, i.e. by having access to large amounts of sense-disambiguated training data in the same language as the test data. Obtaining such large-scale sense-annotated corpora is extremely cumbersome and known to be affected by the *knowledge acquisition bottleneck* (Gale et al., 1992), for which reason solutions that can utilize the training data composed for different languages are of utmost importance.

In this work, we evaluate WSD systems for Hungarian in the cross-lingual and zero-shot setting, as we solely use English sense-annotated data for training, whereas our primary interest is applying this model on Hungarian input texts. We bridge this potential mismatch in the language of input texts during training and test time by relying on multilingual contextualized word representations

which have been shown to yield high quality multilingual representations (Chi et al., 2020; Dufter and Schütze, 2020) that makes them suitable for application in cross-lingual zero-shot settings. We make our source code for reproducing our experiments available at https://github.com/begab/sparsity_makes_sense.

2 Related work

Contextualized word representations, such as CoVE (McCann et al., 2017), ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019), are the most prominent forms of representing the meaning of linguistic units nowadays. Contextualized word representations are *typically* built on the transformer architecture (Vaswani et al., 2017) by employing some kind of masked language modeling objective.

The fact that these models can be trained in a self-supervised manner, i.e. they do not require any explicit labeled data, but raw text only, allows them to be trained on data at an unprecedented scale. As a consequence of being trained on a wide variety of textual utterances, such models have the ability of developing such representations that can capture a wide variety of linguistic phenomena Tenney et al. (2019); Hewitt and Manning (2019); Reif et al. (2019). This is a useful byproduct of these architectures as their training procedures do not explicitly encourage them in becoming able to capture these linguistic phenomena.

Transformer-based language models trained on multilingual texts – without supposing any kind of alignment between the multilingual text passages – have been shown to provide such representations that perform surprisingly well across different languages (K et al., 2020; Chi et al., 2020; Dufter and Schütze, 2020). This property of multilingual transformers opens the possibility of utilizing them in zero-shot settings where the language of a training data do not need to match that of the test set for modeling certain linguistic phenomena.

Recent studies have shown that transformer-based language models that provide contextualized word representations can output extremely valuable inputs to WSD systems even when they are applied in a simple k -nearest neighbor classifier (Loureiro and Jorge, 2019). The application of contextualized word representations with additional sparsity constraints have been recently reported to yield extra improvement for WSD (Berend, 2020a).

These earlier works were, however, focusing on the evaluation of applying contextualized word representations in monolingual WSD settings, i.e. when the sense annotated training corpus is available in the same language relative to the test set. Our work is different from this prior line of research in that we are investigating the performance of these techniques when used in conjunction with multilingual contextualized word representations and evaluate them in the zero-shot setting when the language of the sense annotated training set does not necessarily match the language of the test set.

Most recently, Scarlini et al. (2020) proposed ARES (context-AwaRe Embeddings of Senses) for obtaining sense prototype embeddings that can be used

for WSD in a 1-nearest neighbor fashion similar to (Loureiro and Jorge, 2019). ARES introduces a methodology to exploit useful signal upon the construction of sense embeddings from external knowledge stored in the SyntagNet (Maru et al., 2019). ARES embeddings for all the English WordNet synsets obtained by relying on multilingual BERT were made publicly available by the authors, which makes their utilization possible for languages beside English as well.

3 Experiments

Most successful WSD systems are based on supervision (Raganato et al., 2017). That is, they require (large) sense-annotated training signal in the language of the test sentences. The largest sense annotated training data are in English and use the Princeton WordNet (Fellbaum, 1998) as the basis of sense inventory for disambiguating the distinct senses of the words in their context.

Our evaluation departs from the typical setting, i.e. we rely on English sense-annotated training data for training and evaluate the created WSD model by distinguishing senses of ambiguous words in Hungarian sentences.

3.1 Dataset

We first introduce the English sense-annotated corpora that we used for training our WSD models. We also provide details on the evaluation datasets in both English and Hungarian that we evaluated our models on.

English data We evaluated our approach using the unified WSD evaluation framework (Raganato et al., 2017) that includes the sense-annotated SemCor dataset for training purposes. SemCor Miller et al. (1994) contains 802,443 tokens, out of which more than 28% (226,036) is sense-annotated according to WordNet sensekeys.

We also used the contents of the glosses of the English WordNet synsets and the Princeton WordNet Gloss Corpus (WNGC) as additional sources for constructing our models. WNGC is a sense-annotated version of the WordNet definitions themselves, hence it can be basically used as an extension to the SemCor annotated training set (Vial et al., 2019).

The evaluation framework introduced in (Raganato et al., 2017) also contains five different all-words WSD benchmarks for measuring the performance of WSD systems in English. We also used those for measuring the performance of our WSD models that were based on the contextualized word representations from a multilingual language model. This means that the self-supervised pre-training was conducted over multilingual data, but the creation of our all-words WSD model was both trained and evaluated on English for these experiments. Our evaluation dataset in English consisted of the concatenated test set of the SensEval2 Edmonds and Cotton (2001), SensEval3 Mihalea et al. (2004), SemEval 2007 Task 17 Pradhan et al. (2007), SemEval 2013 Task 12 Navigli et al.

(2013), SemEval 2015 Task 13 Moro and Navigli (2015) shared tasks, comprising of 7253 sense-annotated tokens in total. During our evaluation, we relied on the official scoring script included in the evaluation framework from (Raganato et al., 2017) in our monolingual experiments.

Hungarian test set The dataset we used for our experiments is from (Berend, 2020b), which is a distilled version of the sense-annotated corpus introduced in Vincze et al. (2008). The original dataset contains a collection of documents written in Hungarian that are part of the Hungarian National Corpus (HNC) (Váradí, 2002). The difference between the two datasets is that whereas the former consists of sentences containing ambiguous words, the latter also contains the entire documents for those sentences.

Our corpus that we used for evaluating the performance of our WSD models in Hungarian includes 12,477 sentences, each containing an ambiguous word along with its sense ambiguated label. The sense-annotated dataset contains sense-disambiguated occurrences of 39 different word forms (and their morphologically inflected variants). The 39 distinct word forms were manually disambiguated to one of 200 distinct senses.

3.2 Approach

It was shown recently that by using transformer-based masked language models, such as BERT (Devlin et al., 2019), it becomes possible to build WSD systems by simply obtaining the contextualized embeddings for occurrences of ambiguous words from a sense annotated corpora and performing WSD using nearest neighbor classification for test words (Loureiro and Jorge, 2019). This approach (coined as LMMS by its authors) was, however, both solely trained and evaluated on the unified WSD framework from Raganato et al. (2017), in which cross-linguality did not play a role as both the training and validation corpora were in English.

A recent modification of the LMMS approach proposed the utilization of sparse contextualized word representations and the reliance on the analysis of the sparsity structure of the sense annotated word vectors (Berend, 2020a). We shall refer to this variant of the LMMS approach as S-LMMS throughout the paper, where the prefix is meant to denote that this model variant is based on sparse contextualized word representations.

Performing WSD using the S-LMMS algorithm has been reported to outperform the LMMS strategy significantly when being both trained and evaluated on English, however, it remains a question if the superiority of S-LMMS can be observed in the cross-lingual unsupervised setting as well. Additionally, we explore the effects of using different transformer-based masked language models in our experiments, whereas the original work reported results only on the application of the large cased BERT model.

S-LMMS has two hyperparameters, the dimensionality of the sparse vectors K and the regularization coefficient λ which controls the level of sparsity of the

vectors. We decided to use the same values that was suggested in the original paper, i.e. $K = 3000$ and $\lambda = 0.05$.

3.3 Evaluation

Our evaluation ranges over the investigation of multiple transformer-based multilingual masked language models as inputs to the algorithms we conducted our experiments with, i.e. we relied on the application of multilingual BERT (Devlin et al., 2019), referenced as mBERT hereon, and different versions of the multilingual XLM-Roberta (Conneau et al., 2020) architecture, the XLM-Roberta base and large models.

We used the transformers library (Wolf et al., 2020) to obtain the contextualized multilingual embeddings for our experiments. Even though there exist BERT models specially dedicated to the processing of Hungarian texts, e.g. (Nemeskey, 2020), applying such monolingual models would not suit our setting, since there is a shortage of sense-annotated training data of reasonable size for Hungarian that would be required for training the WSD models. Hence, we mitigate the knowledge acquisition bottleneck of obtaining a high-coverage all-words WSD dataset in Hungarian by using multilingual transformer models and sense-annotated training data in English.

For each model we experiment with, we evaluate the utility of using the contextualized embeddings from the last four layers of the transformer models as well as taking the average of these four layers. The base models consist of 12 layers, whereas the large transformer model has 24 layers.

3.4 Evaluation in English

First, we trained the LMMS and S-LMMS all words WSD models in English and evaluated their utility in the monolingual setting, i.e. on the standard benchmark validation set from (Raganato et al., 2017). Table 1 contains the F1-scores of the different models when relying on contextualized embeddings produced by different multilingual transformer architectures.

Interestingly, the performance of the LMMS models are relatively stable when using multilingual contextualized transformer models of different size as input, however, the Roberta large model has a clear advantage once the contextualized representations it produces are fed into the S-LMMS approach.

Table 1 additionally confirms the findings from (Berend, 2020a), i.e. the S-LMMS variant has a clear advantage over the application of the LMMS model for all the layers and transformer models. There is nonetheless one key methodological difference in our experimental settings compared to those of (Berend, 2020a). Even though the S-LMMS approach and the training/validation data were the same, our work builds on top of multilingual contextualized word representations as input, instead of using the English monolingual BERT large model as input.

The reliance on multilingual models caused some performance loss compared to the best results reported in (Berend, 2020a) (75.7 vs. 78.8), nonetheless the

Layer(s)	LMMS	S-LMMS	Layer(s)	LMMS	S-LMMS	Layer(s)	LMMS	S-LMMS
9	0.698	0.740	9	0.684	0.735	21	0.702	0.757
10	0.702	0.742	10	0.701	0.749	22	0.692	0.753
11	0.706	0.746	11	0.702	0.753	23	0.679	0.749
12	0.699	0.738	12	0.686	0.736	24	0.648	0.728
last-4	0.704	0.743	last-4	0.695	0.748	last-4	0.692	0.754
(a) mBERT			(b) XLM-Roberta-base			(c) XLM-Roberta-large		

Table 1. F-scores obtained over the standard English WSD evaluation benchmark dataset from (Raganato et al., 2017). Results range over the application of various final layers (and their averaging) of different transformer architectures as input to LMMS and S-LMMS.

application of multilingual contextualized transformation is a key component for using the trained model for all-words WSD purposes in the cross-lingual setting.

3.5 Evaluation in Hungarian

Subsequently, we conducted experiments for assessing the quality of the WSD models when applied on Hungarian input texts. When evaluating our models in this cross-lingual setting, we faced the problem that the word senses our models produce are based on the sense inventory of the English WordNet, however, there is no one-to-one correspondence between the synsets of the English WordNet and the senses of the Hungarian sense-disambiguated corpus we used for evaluation.

Our first attempt was create a manual alignment between the senses in the Hungarian dataset and the English WordNet, however, no one-to-one correspondence could be established between the two inventories, as the Hungarian dataset includes such senses that either do not exist in the WordNet or which can correspond to multiple WordNet synsets.

To overcome this issue, we first decide on an evaluation metric that was originally introduced for measuring the performance of clustering techniques. This metric is the V-score (Rosenberg and Hirschberg, 2007), which is similar in nature to the well-known F-score that is intended to be used for the evaluation of classification algorithms. Applying V-score is advantageous, as it can handle situations when we the number of predicted categories and that of the gold standard labels mismatch. This was exactly the situation during our evaluations as we assigned one of the 117,659 English WordNet synsets to the target words that were labeled according to one of the 200 gold standard labels in the dataset. As such, we had 200 gold standard groups of words on the one hand, and as many predicted clusters as many distinct WordNet synsets were assigned to the ambiguous words in the dataset by our algorithms on the other hand.

Just like F-score is the harmonic mean of the precision and the recall of an algorithm, V-score is the harmonic mean of the homogeneity and the complete-

ness scores that are meant to be a respective generalization of the precision and recall scores employed for clustering.

The V-scores between the predicted synset labels from the English WordNet and the gold standard senses according to our evaluation dataset over the 12,477 sense-disambiguated Hungarian words are included in Table 2.

Even though applying V-score for assessing the quality of the implicit clustering based on the most likely synset our models predicted for the sense-annotated Hungarian words is a viable approach, it arguably gives us a too pessimistic view on the true quality of our models. This is partly because our model predicts one of the 117,659 different English WordNet synsets, whereas there were only 200 distinct senses distinguished in the sense-annotated corpus we based our evaluation on.

As such, our model often ended up producing near-misses, such as the (co-)hypernym/hyponym of the correct senses of a word. Additionally, due to the mismatch of the employed labels, there were certain cases when there would be no exact match in the English WordNet for some Hungarian sense labels.

Layer(s)	LMMS	S-LMMS	Layer(s)	LMMS	S-LMMS	Layer(s)	LMMS	S-LMMS
9	0.216	0.209	9	0.242	0.236	21	0.262	0.277
10	0.218	0.220	10	0.259	0.261	22	0.246	0.272
11	0.197	0.219	11	0.255	0.262	23	0.229	0.262
12	0.196	0.207	12	0.239	0.233	24	0.224	0.227
last-4	0.220	0.220	last-4	0.254	0.258	last-4	0.253	0.274
(a) mBERT			(b) XLM-Roberta-base			(c) XLM-Roberta-large		

Table 2. V-scores averaged over the ambiguous word forms obtained from the Hungarian WSD evaluation benchmark dataset. Results range over the application of various final layers (and their averaging) of different transformer architectures as input to LMMS and S-LMMS.

Due to the above reasons, we assessed the quality of our cross-lingual WSD predictions in an alternative manner for our subsequent experiment. Instead of expecting the models to find a good one-to-one mapping between the English synsets and the set of sense labels included in our Hungarian evaluation set (which does not even exist in the first place for certain senses by the design of the sense labels of the two different sense inventories), we quantified the extent to which the ordered list of the English synsets that our models assigned to the Hungarian ambiguous words are similar for those words that received the same sense label in our evaluation dataset in Hungarian.

For each ambiguous Hungarian word in our test set, we determined the top-15 English synsets our model assigned to them. As a subsequent step, we calculated the similarity between all pairs of ambiguous words based on the ordered list of most likely English synsets assigned to the word occurrences.

Finally, we determined the nearest neighbor of each ambiguous word according to their similarity and quantified the relative number of times the ground truth sense label of nearest neighbor words were identical. As such, these results can be viewed as the performance of a 1-nearest-neighbor classifier that determines the proximity of word occurrences based on the ordered list of most likely synsets that our cross-lingual model assigns to them.

The top-15 most likely synsets assigned to a pair of ambiguous words are non-conjoint, meaning that they can differ to any extent. Indeed, in the most extreme case, there could be no overlap at all between the ranked lists of synsets. To this end, we measured the similarities between the top-ranked synsets for a pair of words using the pairwise ranking-biased overlap (RBO) (Webber et al., 2010) score, which (among others) has the favorable property of being capable of measuring the similarity between non-conjoint ordered lists.

These results are included in Table 3. Similar to the earlier results in Table 1 and Table 2, i.e. the utilization of S-LMMS with input representations originating from the 21th layer of the large XLM-Roberta model provided the best results according to the nearest-neighbor accuracy metric.

Layer(s)	LMMS	S-LMMS	Layer(s)	LMMS	S-LMMS	Layer(s)	LMMS	S-LMMS
9	0.756	0.740	9	0.792	0.795	21	0.828	0.830
10	0.756	0.752	10	0.811	0.815	22	0.818	0.821
11	0.740	0.742	11	0.808	0.815	23	0.809	0.814
12	0.730	0.716	12	0.796	0.795	24	0.791	0.777
last-4	0.761	0.757	last-4	0.801	0.805	last-4	0.823	0.824
(a) mBERT			(b) XLM-Roberta-base			(c) XLM-Roberta-large		

Table 3. Nearest neighbor accuracy averaged over the ambiguous word forms obtained from the Hungarian WSD evaluation benchmark dataset. Results range over the application of various final layers (and their averaging) of different transformer architectures as input to LMMS and S-LMMS.

3.6 Comparison with ARES embeddings

As mentioned in Section 2, ARES embeddings (Scarlina et al., 2020) can be utilized for WSD similar to LMMS (Loureiro and Jorge, 2019) by performing a 1-nearest neighbor search between the contextualized embedding of some word and the pre-calculated sense embeddings. The important difference between ARES and LMMS is that ARES also uses a semi-supervised approach to improve the sense embeddings obtained for those WordNet synsets with no/few occurrences in the sense-annotated training corpus, e.g. SemCor. A key important property of the ARES embeddings from our point is that such a variant that builds upon the contextualized representations of the last 4 layers of mBERT is made available.

In Table 4, we compare our previously reported results to those obtained with ARES. For better comparability, we chose to compare the results of the (S-)LMMS approaches that rely on the same contextualized representations as ARES, i.e. the averaged outputs of the last 4 layers of mBERT. S-LMMS performed clearly better on the English test set, whereas our evaluation on the Hungarian data shows a mixed, but comparable behavior for the different approaches.

ARES LMMS S-LMMS				ARES LMMS S-LMMS			
Accuracy	0.713	0.704	0.743	V-score	0.208	0.220	0.220
				Accuracy	0.765	0.761	0.757

(a) Evaluation on English.

(b) Evaluation on Hungarian.

Table 4. Comparing the application of ARES embeddings and our models using the last 4 layers of mBERT embeddings.

4 Conclusions

Transformer-based language models are known to provide extremely valuable input for WSD. Our paper investigated the possibility of exploiting the sense-annotated training corpus of some resource-rich source language, and utilizing the trained WSD model on some distinct target language by relying on the cross-lingual interoperability of the multilingual contextualized word representation produced by mBERT and XLM-Roberta. The proposed approach showed a promising solution for overcoming the need of a large high-coverage sense-annotated training corpus.

Acknowledgments

The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program and the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008).

The author is grateful for the fruitful discussions on the topic of word sense disambiguation with Ádám Szórád whose research was funded by the project ”Integrated program for training new generation of scientists in the fields of computer science”, no EFOP-3.6.3-VEKOP-16-2017-0002, supported by the EU and co-funded by the European Social Fund.

Bibliography

- Berend, G.: Sparsity makes sense: Word sense disambiguation using sparse contextualized word representations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8498–8508. Association for Computational Linguistics, Online (Nov 2020a), <https://www.aclweb.org/anthology/2020.emnlp-main.683>
- Berend, G.: Word sense disambiguation for hungarian using transformers. In: Berend, G., Gosztolya, G., Vincze, V. (eds.) XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020). p. 3–13. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2020b)
- Chi, E.A., Hewitt, J., Manning, C.D.: Finding universal grammatical relations in multilingual BERT. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5564–5577. Association for Computational Linguistics, Online (Jul 2020), <https://www.aclweb.org/anthology/2020.acl-main.493>
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020), <https://www.aclweb.org/anthology/2020.acl-main.747>
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), <https://www.aclweb.org/anthology/N19-1423>
- Dufter, P., Schütze, H.: Identifying elements essential for BERT’s multilinguality. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4423–4437. Association for Computational Linguistics, Online (Nov 2020), <https://www.aclweb.org/anthology/2020.emnlp-main.358>
- Edmonds, P., Cotton, S.: SENSEVAL-2: Overview. In: The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems. pp. 1–5. SENSEVAL ’01, Association for Computational Linguistics, Stroudsburg, PA, USA (2001), <http://dl.acm.org/citation.cfm?id=2387364.2387365>
- Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
- Gale, W.A., Church, K.W., Yarowsky, D.: A method for disambiguating word senses in a large corpus. Computers and the Humanities 26(5), 415–439 (Dec 1992), <https://doi.org/10.1007/BF00136984>
- Hewitt, J., Manning, C.D.: A structural probe for finding syntax in word representations. In: Proceedings of the 2019 Conference of the North Ameri-

- can Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4129–4138. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), <https://www.aclweb.org/anthology/N19-1419>
- K, K., Wang, Z., Mayhew, S., Roth, D.: Cross-lingual ability of multilingual BERT: An empirical study. In: International Conference on Learning Representations (2020), <https://openreview.net/forum?id=HJeT3yrtDr>
- Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In: Proceedings of the 5th Annual International Conference on Systems Documentation. pp. 24–26. SIGDOC '86, ACM, New York, NY, USA (1986), <http://doi.acm.org/10.1145/318723.318728>
- Loureiro, D., Jorge, A.: Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5682–5691. Association for Computational Linguistics, Florence, Italy (Jul 2019), <https://www.aclweb.org/anthology/P19-1569>
- Maru, M., Scozzafava, F., Martelli, F., Navigli, R.: SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3534–3540. Association for Computational Linguistics, Hong Kong, China (Nov 2019), <https://www.aclweb.org/anthology/D19-1359>
- McCann, B., Bradbury, J., Xiong, C., Socher, R.: Learned in translation: Contextualized word vectors. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 6294–6305. Curran Associates, Inc. (2017), <http://papers.nips.cc/paper/7209-learned-in-translation-contextualized-word-vectors.pdf>
- Mihalcea, R., Chklovski, T., Kilgariff, A.: The senseval-3 english lexical sample task. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. pp. 25–28. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), <https://www.aclweb.org/anthology/W04-0807>
- Miller, G.A., Chodorow, M., Landes, S., Leacock, C., Thomas, R.G.: Using a semantic concordance for sense identification. In: HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994 (1994), <https://www.aclweb.org/anthology/H94-1046>
- Moro, A., Navigli, R.: SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 288–297. Association for Computational Linguistics, Denver, Colorado (Jun 2015), <https://www.aclweb.org/anthology/S15-2049>
- Navigli, R.: Word sense disambiguation: A survey. ACM Comput. Surv. 41(2) (Feb 2009), <https://doi.org/10.1145/1459352.1459355>

- Navigli, R., Jurgens, D., Vannella, D.: SemEval-2013 task 12: Multilingual word sense disambiguation. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). pp. 222–231. Association for Computational Linguistics, Atlanta, Georgia, USA (Jun 2013), <https://www.aclweb.org/anthology/S13-2040>
- Nemeskey, D.M.: Natural Language Processing Methods for Language Modeling. Ph.D. thesis, Eötvös Loránd University (2020)
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018), <https://www.aclweb.org/anthology/N18-1202>
- Pradhan, S.S., Loper, E., Dligach, D., Palmer, M.: Semeval-2007 task 17: English lexical sample, srl and all words. In: Proceedings of the 4th International Workshop on Semantic Evaluations. pp. 87–92. SemEval '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), <http://dl.acm.org/citation.cfm?id=1621474.1621490>
- Raganato, A., Camacho-Collados, J., Navigli, R.: Word sense disambiguation: A unified evaluation framework and empirical comparison. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 99–110. Association for Computational Linguistics, Valencia, Spain (Apr 2017), <https://www.aclweb.org/anthology/E17-1010>
- Reif, E., Yuan, A., Wattenberg, M., Viegas, F.B., Coenen, A., Pearce, A., Kim, B.: Visualizing and measuring the geometry of bert. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32, pp. 8594–8603. Curran Associates, Inc. (2019), <https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdf>
- Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 410–420. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), <https://www.aclweb.org/anthology/D07-1043>
- Scarlini, B., Pasini, T., Navigli, R.: With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3528–3539. Association for Computational Linguistics, Online (Nov 2020), <https://www.aclweb.org/anthology/2020.emnlp-main.285>
- Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of the 57th Annual Meeting of the Association for Computa-

- tional Linguistics. pp. 4593–4601. Association for Computational Linguistics, Florence, Italy (Jul 2019), <https://www.aclweb.org/anthology/P19-1452>
- Váradi, T.: The Hungarian national corpus. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain (May 2002), <http://www.lrec-conf.org/proceedings/lrec2002/pdf/217.pdf>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017), <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- Vial, L., Lecouteux, B., Schwab, D.: Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation. In: Global Wordnet Conference. Wroclaw, Poland (2019), <https://hal.archives-ouvertes.fr/hal-02131872>
- Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J.: Hungarian word-sense disambiguated corpus. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (May 2008)
- Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. 28(4) (Nov 2010), <https://doi.org/10.1145/1852102.1852106>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- Zhong, Z., Ng, H.T.: Word sense disambiguation improves information retrieval. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 273–282. Association for Computational Linguistics, Jeju Island, Korea (Jul 2012), <https://www.aclweb.org/anthology/P12-1029>

Analysing the semantic content of static Hungarian embedding spaces

Tamás Ficsor¹, Gábor Berend^{1,2}

¹ Institute of Informatics, University of Szeged, Hungary

² MTA-SZTE Research Group on Artificial Intelligence
`{ficsor,berend}@inf.u-szeged.hu`

Abstract. Word embeddings can encode semantic features and have achieved many recent successes in solving NLP tasks. Although word embeddings have high success on several downstream tasks, there is no trivial approach to extract lexical information from them. We propose a transformation that amplifies desired semantic features in the basis of the embedding space. We generate these semantic features by a distant supervised approach, to make them applicable for Hungarian embedding spaces. We propose the Hellinger distance in order to perform a transformation to an interpretable embedding space. Furthermore, we extend our research to sparse word representations as well, since sparse representations are considered to be highly interpretable.

Keywords: Interpretability, Semantic Transformation, Word Embeddings

1 Introduction

Continuous vectorial word representations are routinely employed as the inputs of various NLP models such as named entity recognition (Seok et al., 2016), part of speech tagging (Abka, 2016), question answering (Shen et al., 2015), text summarization (Mohd et al., 2020), dialog systems (Forgues et al., 2014) and machine translation (Zou et al., 2013).

Static word representations acquire their lexical knowledge from local or global contexts. GloVe (Pennington et al., 2014a) uses global co-occurrence statistics to determine a word’s representation in the continuous space, whereas Mikolov et al. (2013) proposed a predictive model for predicting target words from their contexts. Furthermore, Bojanowski et al. (2017) presented a training technique of word representations where sub-word information is in the form of character n -grams are also considered. The outputs of these word embedding algorithms are able to encode semantic relations between words (Pennington et al., 2014a; Nugaliyadde et al., 2019). This can be present on word-level – such as similarity in meaning, word analogy, antonymic relation – or word embeddings can be utilized to produce sentence-level embeddings, which shows that word vectors still carry intra-sentence information (Kenter and de Rijke, 2015).

Despite the successes of word embeddings on semantics related tasks, we have no direct knowledge of the human-interpretable information contents of dense

dimensions. Utilizing human-interpretable features as prior information could lead to performance gain in various NLP tasks. Identifying and understanding the dense representation in each dimension can be cumbersome for humans. To alleviate this problem, we propose a transformation where we map existing word representations into a more interpretable space, where each dimension is supposed to be responsible for encoding semantic information from a predefined set of semantic inventory. There are various ways to form groups of semantic classes by forming semantically coherent groups of words. In this work, we shall rely on ConceptNet (Speer et al., 2016) to do so.

We measure the information contents of each dimension in the original embedding space towards a predefined set of human interpretable concepts. Our approach is inspired by Şenel et al. (2018) which utilized the Bhattacharyya distance for the aforementioned purpose. In this work, we also evaluate a close variant of the Bhattacharyya distance, the Hellinger distance for transforming word representations in a way that the individual dimensions have a more transparent interpretation.

Feature norming studies have revealed that humans usually tend to describe the properties of objects and concepts with a limited number of sparse features (McRae et al., 2005). This kind of sparse representation became a major part of natural language processing since we can see the resemblance between sparse features and human feature descriptions. Hence, we additionally explore the effects of applying sparse word representations as an input to our algorithm which makes the semantic information stored along the individual dimensions more explicit. We published our work on GitHub for interpretable word vector generation: https://github.com/ficstamas/word_embedding_interpretability, and shared the code for semantic category generation as well, alongside with the used semantic categories: https://github.com/ficstamas/multilingual_semantic_categories.

2 Related Work

Turian et al. (2010) was one of the first providing a comparison of several word embedding methods and showed that incorporating them into established NLP pipelines can also boost their performance. word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014b) and Fasttext (Bojanowski et al., 2017) methods are well known models for obtaining context-insensitive (or static) word representations. These methods generate static word vectors, i.e. every word form gets assigned a single vector that applies to all of its occurrences and senses.

The intuition behind sparse vectors is related to the way humans interpret features, which was shown in various feature norming studies (Garrard et al., 2001; McRae et al., 2005). Additionally, generating sparse features (Kazama and Tsujii, 2003; Friedman et al., 2008; Mairal et al., 2009) has proved to be useful in several areas of NLP, including POS tagging (Ganchev et al., 2010), text classification (Yogatama and Smith, 2014) and dependency parsing (Martins et al., 2011). Berend (2017) also showed that sparse representations can outperform their

	Ours	SemCat	HyperLex
Number of Categories	91	110	1399
Number of Unique Words	2760	6559	1752
Average Word Count per Category	68	91	2
Standard Deviation of Word Counts	52	56	3

Table 1. Basic statistics about the semantic categories.

dense counterparts in certain NLP tasks, such as NER, or POS tagging. Murphy et al. (2012) proposed Non-Negative Sparse Embedding to learn interpretable sparse word vectors, Park et al. (2017) showed a rotation based method and Subramanian et al. (2017) suggested an approach using a denoising k-sparse auto-encoder to generate interpretable sparse word representations. Balogh et al. (2019) made prior research about the semantic overlap of the generated vectors with a human commonsense knowledgebase and found that substantial semantic content is captured by the bases of sparse embedding space.

Şenel et al. (2018) showed a method where they measured the interpretability of the dense GloVe embedding space, and later showed a method to manipulate and improve the interpretability of a given static word representation (Şenel et al., 2020).

Our proposed approach also relates to the application of the Hellinger distance, which has been used in NLP for constructing word embeddings Lebrete and Collobert (2014). Note that the way we apply the Hellinger distance differs from prior work in that we use it for amplifying the interpretability of contextual word representations, whereas the Hellinger distance served as the basis for constructing (static) embeddings in earlier work.

3 Data

3.1 Semantic Categories

Amplifying and understanding the semantic contents from word embedding spaces is the main objective of this study. To provide meaningful interpretation to each dimension, we rely on the base concept of distributional semantics (Harris, 1954; Boleda, 2020). In order to investigate the underlying semantic properties of word embeddings, we have to define some kind of semantic categories that represent the semantic properties of words. These semantic properties can represent any arbitrary relation which makes sense from a human perspective, for example, words such as *"red"*, *"green"*, and *"yellow"* can be grouped under the **"color"** semantic category which represents a hypernym-hyponym relation, but they can be found among *"traffic"* related terms as well. Another example is **"car"** semantic category which is in meronymy relation with words such as *"engine"*, *"wheels"* and *"crankcase"*.

Previous similar linguistic resources that contain semantic categorization of words include HyperLex (Véronis, 2004) and SemCat (Şenel et al., 2018). A

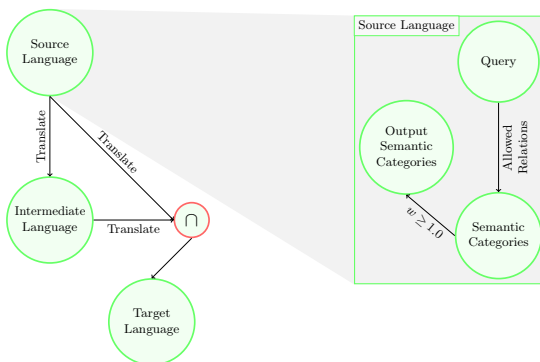


Fig. 1. Generation of semantic categories with the help of allowed relations from ConceptNet, where the Query represents the root concept, and w denotes the weight of the relation.

major problem with them from the standpoint of applicability is that these datasets are restricted to English, so they can not be utilized in multilingual scenarios. From an informational standpoint, HyperLex with a low average and standard deviation category sizes also raises concerns. In order to extend it to the Hungarian language as well, we used the semantic category names from SemCat and defined relations on a category-by-category base manually. We relied on a subset of relations from ConceptNet (Speer et al., 2016). To obtain higher quality semantic categories, we introduced an intermediate language that works as a validation to reduce undesired translations. The whole process can be followed in Figure 1.

First, we generate the semantic categories from the source language by the allowed relations and restricted the inclusion of words by the weight of the relation. Semantic category names from SemCat were used as the input (Query) and the weight of each relation is originated from ConceptNet. Then we translate the semantic categories to the target language directly and through the intermediate language to the target language, where we kept the intersection of the two results. It is recommended to rely on one of the **core** languages defined in ConceptNet as Source and Intermediate language. Using ConceptNet for inducing the semantic categories for our experiments makes it easy to extend our experiments later for additional languages beyond Hungarian. We present some basic statistics about the mentioned semantic categories in Table 1. This kind of distant supervised generation (Mintz et al., 2009) can produce large number of data easily but it carries the possibility that the generated data is noisy.

3.2 Word Embeddings

We conducted our experiments on 3 embedding spaces trained using the Fast-text algorithm (Bojanowski et al., 2017). The 3 embedding spaces that we relied on were the Hungarian Fasttext (Fasttext HU) embeddings pre-trained on

Wikipedia³, its aligned variant⁴ (Fasttext Aligned) that was created using the RCSLS criteria (Joulin et al., 2018) with the objective to bring Hungarian embeddings closer to semantically similar English embeddings and the Szeged Word Vectors (Szeged WV) (Szántó et al., 2017) which is based on the concatenation of multiple Hungarian corpora.

We limited the word embeddings to their 50,000 most frequent tokens and evaluated every experiment with this subset of all vectors. The vocabulary of the Fasttext HU and Fasttext Aligned embeddings are identical, however, it is important to emphasize that the Szeged WV overlap with the vocabulary of these embedding spaces on less than half of the word forms, i.e. 22,112 words. Furthermore, Szeged WV uses a cased vocabulary, unlike the Fasttext embeddings. In the case of Fasttext, the vocabulary of the embedding and our semantic categories overlaps in 1848 unique words. For the Szeged WV, it only overlaps with 1595 unique words.

Our approach can evaluate other embedding types as well. So due to the fact that sparse embeddings are deemed to be more interpretable compared to their dense counterparts, we also produced sparse static word representations by applying dictionary learning for sparse coding (Mairal et al., 2009) (DLSC) on the dense representation. For obtaining the sparse word representations of dense static embedding space \mathcal{E} , we solved the optimization problem

$$\min_{\alpha, D} \frac{1}{2} \|\mathcal{E} - \alpha D\|_F^2 + \lambda \|\alpha\|_1,$$

that is, our goal is to decompose $\mathcal{E} \in \mathbb{R}^{v \times d}$ into the product of a dictionary matrix $D \in \mathbb{R}^{k \times d}$ and a matrix of sparse coefficients $\alpha \in \mathbb{R}^{v \times k}$ with a sparsity-inducing ℓ_1 penalty on the elements of α . Furthermore, v denotes the size of the vocabulary, d represents the dimensionality of the original embedding space, and k is the number of basis vectors.

We obtained different sparse embedding space by modifying the hyperparameters of the algorithm. So we evaluated it with $\lambda \in \{0.05, 0.1, 0.2\}$ regularization and $k \in \{1000, 1500, 2000\}$ basis vectors.

4 Our Approach

4.1 Semantic Decomposition

The foundation of our approach is to measure the encoding of semantic information in the basis of pre-trained static word embeddings. In order to quantify the semantic information, we have to observe the joint behavior of similarities in semantic distributions. This approach is feasible due to distributional semantics (Boleda, 2020), which states that similarity in meaning results in similarity in linguistic distribution (Harris, 1954). This behavior can be observed from the

³ <https://dl.fbaipublicfiles.com/fasttext/vectors-wiki/wiki.hu.vec>

⁴ <https://dl.fbaipublicfiles.com/fasttext/vectors-aligned/wiki.hu.align.vec>

fact that static word representations are trained on co-occurrence information of word tokens. So if we are able to measure the dissimilarity between a distribution that represents a semantic information and the distribution of space (which is the complementary distribution of semantic information) then we can give a transformation that is going to explicitly express the semantic categories in each dimension.

In other words, the coefficients of a dimension form a distribution $\mathcal{R} \in \mathbb{R}^v$. The desired semantic information we try to express is denoted as $\mathcal{P} \subseteq \mathcal{R}$. For example, \mathcal{P} describes the "wave" semantic information, then words related to that term should occur in a similar context, such as "rising", "golden", or "lacy" in " _ waves". So by expressing how far this distribution is from the distribution of a dimension, then we can see how significant is the dimension about the semantic information. The certainty of such a dimension about the desired semantic information can be formulated as $D(\mathcal{P}, \overline{\mathcal{P}})$. If this distance is low then it means that the information gain would be really low because the two distributions are nearly homogeneous. Analogously, if the distance is high then we can rely on that dimension with higher certainty. So the distance expresses the certainty we have in each dimension about the semantic information.

In order to express the certainty in a dimension, first, we have to separate the coefficients in a dimension to represent the previously defined distributions. As a reminder, we denoted the embedding space with \mathcal{E} , then we denote the defined semantic categories as \mathcal{S} . So we can define function $f : x \rightarrow \mathcal{E}$ which returns the representation of word token x , and function $S : x \rightarrow \mathcal{S}$ which maps word token x to its corresponding semantic category. Then we can separate the coefficients along the i th dimension and j th semantic category as

$$P_{ij} = \left\{ f(x)^{(i)} \mid f(x) \in \mathcal{E}, S(x) \in \mathcal{S}^{(j)} \right\}$$

and similarly

$$Q_{ij} = \left\{ f(x)^{(i)} \mid f(x) \in \mathcal{E}, S(x) \notin \mathcal{S}^{(j)} \right\},$$

where P_{ij} represents the distribution of a particular semantic category in a dimension (in-category words) and Q_{ij} ($= \overline{P_{ij}}$) represents the distribution of the rest of the dimension (out-of-category words).

4.2 Measuring dissimilarity

To measure the dissimilarity, hence observe the certainty of semantic categories in each dimension we define two distances. We apply Bhattacharyya distance as a baseline from Şenel et al. (2018) and Hellinger distance as an alternative improvement. Both distances can be expressed by Bhattacharyya coefficient (or fidelity coefficient) as

$$D_B(p, q) = -\ln \int_{-\infty}^{\infty} \sqrt{p(x)q(x)} dx \quad D_H(p, q) = \sqrt{1 - \int_{-\infty}^{\infty} \sqrt{p(x)q(x)} dx},$$

where the integrand expresses the fidelity coefficient. The important differences between the two types of distances are that

- Hellinger distance is a bounded metric that eases the interpretation of values when the fidelity is close to 0,
- Hellinger distance accumulates small distributional differences better which means if the fidelity is close to 1, it can still enhance potentially significant information.

To maintain consistency, comparability and a baseline, we define Bhattacharyya distance as Şenel et al. (2018), and Hellinger distance by their closed forms which assumes normality of the investigated distributions. Under the normality assumption, the Bhattacharyya distance can be expressed as

$$D_B(P_{i,j}, Q_{i,j}) = \frac{1}{4} \ln \left(\frac{1}{4} \left(\frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2 \right) \right) + \frac{1}{4} \left(\frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2} \right), \quad (1)$$

and Hellinger distance can be formulated as

$$D_H(P_{i,j}, Q_{i,j}) = \sqrt{1 - \sqrt{\frac{2\sigma_p\sigma_q}{\sigma_p^2 + \sigma_q^2} e^{-\frac{1}{4} \cdot \frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2}}}}, \quad (2)$$

where σ denotes the standard deviation and μ denotes the mean of $P_{i,j}$ and $Q_{i,j}$ respectively, assuming that $P_{i,j} \sim \mathcal{N}(\sigma_p, \mu_p)$ and $Q_{i,j} \sim \mathcal{N}(\sigma_q, \mu_q)$. We then define $\mathcal{W}_D \in \mathbb{R}^{d \times |S|}$ that contains the distances of semantic category-dimension pairs, i.e. $\mathcal{W}_D(i, j) = D(P_{ij}, Q_{ij})$, with D denoting either of the Bhattacharyya or Hellinger distances.

4.3 Interpretable Word Vector Generation

In order to obtain interpretable word vectors, we have to first refine the quality of transformation. It is highly possible that our semantic category dataset is imbalanced and/or during the pre-training process we do not have enough information about a word token. So we should reduce the bias of dominant semantic categories which can be obtained by performing ℓ_1 normalization on \mathcal{W}_D in such a manner that the values corresponding to each semantic category sum up to 1. We shall denote the transformation matrix that we derive in such a manner as \mathcal{W}_{ND} .

Another problem which occurs in embedding spaces is that semantic information can be encoded in both positive and negative direction relative to the mean, hence we should adjust the orientation of these vectors in certain bases in order to couple semantic categories in their corresponding bases and segregate them from others in other bases. We determine the directions from the sign of difference between the mean of the original distributions, thus we can obtain \mathcal{W}_{NSD} as

$$\mathcal{W}_{NSD}(i, j) = \text{sign}(\Delta_{ij}) \cdot \mathcal{W}_{ND}(i, j),$$

where $\Delta_{ij} = \mu_{p_{ij}} - \mu_{q_{ij}}$ and sign is the signum function.

We also standardize \mathcal{E} in order to avoid multicollinear issues, thus we can yield higher quality word vectors. We denote the standardized embedding space

by \mathcal{E}_S . As a final step, we obtain our interpretable representations $\mathcal{I} \in \mathbb{R}^{v \times |S|}$ as the product of \mathcal{E}_S and \mathcal{W}_{NSD} .

5 Evaluation methods

5.1 Word Retrieval Test

We are concerned about the accuracy of our model, to know how well it behaves on unknown data. In \mathcal{W}_D we can see the semantic distribution of the dimensions and in \mathcal{I} each column should represent a semantic category. So each dimension in \mathcal{I} should ideally represent a semantic category from the semantic categories.

In order to measure the semantic quality of \mathcal{I} , we used 60% of the words from each semantic category for training and 40% for evaluation. By relying on the training set, we calculate the distance matrix \mathcal{W}_D from the embedding space, using any arbitrary distance we defined earlier. We also experiment with a pruned version of \mathcal{W}_D by keeping the highest \mathcal{K} coefficients for each semantic category and setting the rest to 0, and denoting it as \mathcal{W}_D^S . We do that, so we can inspect the importance of the strongest encoding dimensions. Then by employing \mathcal{W}_D^S instead of \mathcal{W}_D , we do everything in the same way as we defined earlier.

We use the validation set and see whether the words of a semantic category are seen among the top n , $3n$ or $5n$ words in the corresponding dimension in \mathcal{I}_S , where n is the number of the words in the validation set varying across the semantic categories. The final accuracy is calculated as the weighted mean of the accuracy of the dimensions, where the weight is the number of words in each category for the corresponding dimension.

5.2 Interpretability

In order to measure the interpretability of the semantic space, we use a functionally-grounded evaluation method (Doshi-Velez and Kim, 2017), which means it does not involve humans in the process of quantification. Furthermore, we use continuous values to express the level of interpretability (Murdoch et al., 2019).

The metric we rely on is an adaptation of the one proposed in (Şenel et al., 2018). We ought to have a metric that is independent from the dimensionality of the embedding space, so models with different number of dimensions can be compared more meaningfully.

$$IS_{i,j}^+ = \frac{|S_j \cap V_i^+(\beta \times n_j)|}{n_j} \quad (3) \quad IS_{i,j}^- = \frac{|S_j \cap V_i^-(\beta \times n_j)|}{n_j} \quad (4)$$

Eqn. (3) and (4) define the interpretability score for the positive and negative directions, respectively. In both equation i represents the dimension ($i \in \{1, 2, 3, \dots, d\}$, where d is the number of dimensions of the embedding space) and j the semantic categories ($j \in \{1, 2, 3, \dots, c\}$, where c is the number of the semantic categories). S_j represents the set of words belonging to the j th semantic category, n_j the number of words in that semantic category. V_i^+ and V_i^- gives us the top and bottom words selected by the magnitude of their coordinates

β	Hellinger			Bhattacharyya		
	1	5	10	1	5	10
Fasttext HU	22.00	38.43	46.87	21.29	38.80	47.01
Fasttext Aligned	26.81	43.71	51.26	25.92	43.45	51.22
Szeged WV	16.34	31.71	40.04	15.69	31.50	39.91

Table 2. Interpretability of Hungarian Fasttext, Aligned Fasttext and Szeged WV with different β relaxation and applied distance.

respectively along the i th dimension. $\beta \times n_j$ is the number of words selected from the top and bottom words, hence $\beta \in \mathbb{N}^+$ is the relaxation coefficient, as it controls how strict we measure the interpretability. As the interpretability of a dimension-category pair, we take the maximum of the positive and negative direction according to

$$IS_{i,j} = \max \{ IS_{i,j}^+, IS_{i,j}^- \}. \quad (5)$$

Once we have the overall interpretability ($IS_{i,j}$), we calculate the categorical interpretability according to Eqn. (6). Şenel et al. (2018) took a different approach of taking the average of the maximum values over all the categories, however, this could easily overestimate the true interpretability of the embedding space.

In order to avoid the overestimation of the interpretability of the embedding space, we calculate Eqn. (6), where we have a condition on the selected i dimension which is defined by Eqn. (7). It chooses the highest encoding dimension according to \mathcal{W}_D (distance matrix of the examined space) which ensures that we obtain the interpretability score from the most likely encoding dimension. This method is more suitable to obtain the interpretability scores, because it relies on the distribution of the semantic categories, instead of the interpretability score equally sampled from each dimension.

$$IS_j = IS_{i_j^*,j} \times 100 \quad (6) \quad i_j^* = \arg \max_{i'} \mathcal{W}_D(i', j) \quad (7)$$

Finally, we define the overall interpretability of the embedding space by taking the average of the interpretability scores across the semantic categories, $IS = \frac{1}{c} \sum_{j=1}^c IS_j$, where c is the number of categories.

6 Results

6.1 Dense Representations

We transformed all 3 embedding spaces to their interpretable representations and measured the effectiveness of the encoding by the interpretability score which can be seen in Table 2. Furthermore, we measured the generalisability of the transformation with word retrieval test which is presented in Figure 2. These types of evaluations are better observed jointly because they represent a different aspect of the embedding space but we can not make any conclusion without each other.

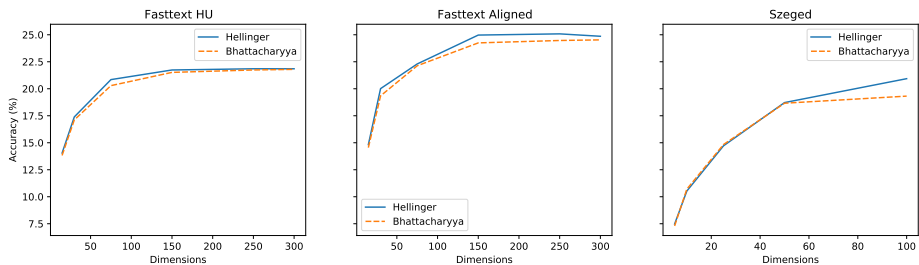


Fig. 2. The results of word retrieval tests with a relaxed size of retrieved words, where the dimensions represent the \mathcal{K} kept coefficient from \mathcal{W}_D .

Fasttext HU				Fasttext Aligned				Szeged WV			
λ	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2		
Hellinger distance											
$k = 1000$	58.11	43.21	19.33	60.13	47.58	24.25	58.88	53.85	33.82		
$k = 1500$	64.49	49.24	23.82	65.50	52.20	28.44	65.03	60.94	38.14		
$k = 2000$	68.29	52.53	26.98	68.79	57.05	30.63	67.65	64.08	42.22		
Bhattacharyya distance											
$k = 1000$	53.20	33.98	18.72	55.54	37.88	22.08	56.13	45.52	27.79		
$k = 1500$	57.77	36.33	21.59	59.91	39.54	24.61	62.85	50.53	30.77		
$k = 2000$	60.82	39.03	24.43	62.99	42.26	26.43	64.45	52.18	33.12		

Table 3. The effects of relying on sparse static word representation with different hyperparameters for regularization coefficient (λ) and number of basis vectors (k). Interpretability scores represented at $\beta = 1$ relaxation.

We can immediately spot the dominant performance on both evaluation methods by the aligned Fasttext word vectors. It can indicate that either the alignment could carry extra semantic knowledge or the English Wikipedia corpus is a higher quality. Szeged WV seems to be the worst-performing model according to interpretability, but it is not necessarily the case because it has a third of the number of dimensions than the Fasttext models, and differ in overlap of words in the vocabulary. In Figure 2 we can also see that it has just enough dimensions (maybe it could utilize a little bit more). This can be seen by observing the accuracy of the embedding spaces. The accuracy has not peaked before relying on all 100 of the dimensions, unlike Fasttext HU which peaks between 150 and 250 dimensions. Furthermore, it does not have a plateau-like effect where we yield little to no improvement. But these observations only apply from the standpoint of our semantic categories, not in a general manner.

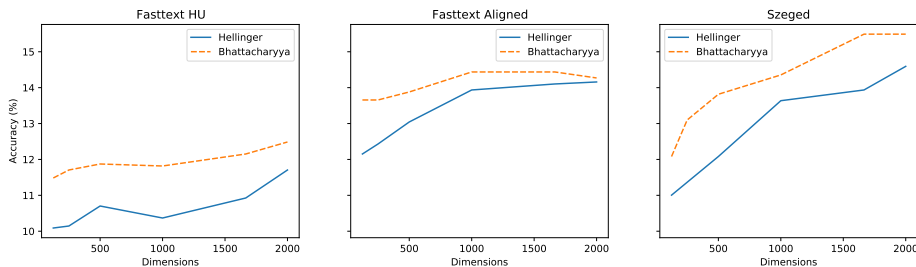


Fig. 3. The results of word retrieval tests on sparse representations ($\lambda = 0.05$ and $k = 2000$), where the dimensions represent the \mathcal{K} kept coefficient from \mathcal{W}_D .

6.2 Sparse Representations

If we closely inspect Eqn. (1) and (2), we can see that division errors occur when σ_p or σ_q equals 0. When the standard deviation for P or Q would be 0, we replace it by $\sqrt{10^{-5}}$ instead.

We evaluated our experiments with different hyperparameters for sparse vector generation as we can see in Table 3 when using the $\beta = 1$ relaxation. We can conclude that increasing the level of sparsity does not benefit the interpretability. On the other hand, changing the number of basis vectors has a beneficial impact. We can see that sparse representation amplifies the semantic information on each basis, since the interpretability of these embedding spaces improved by 2-3 times.

Figure 3 demonstrates the results of the word retrieval test when using sparse representations obtained when setting $\lambda = 0.05$ and $k = 2000$. We can see that the generalisability of the model is decreased overall, and we should rely on more \mathcal{K} none zero coefficients to extract the semantic information. This could be the cause of high level of noise is present in our semantic categories.

6.3 Semantic Decomposition

We can see the semantic decomposition of the word "ember" on Figure 4. In the first row, we represent the dense and in the second we represent the sparse embedding spaces. We expect that in this case for the "ember" word, semantic categories that contain this word are among the highest coefficients. We can see that, after we obtained the sparse representations for Fasttext, and transformed them the semantic decomposition shows an identical representation even though their scores are different.

7 Conclusion

We evaluated the transformation of non-contextual embedding spaces into a more interpretable one, which can be used to analyze the semantic distribution which can have a potential application in knowledge base completion. We investigated

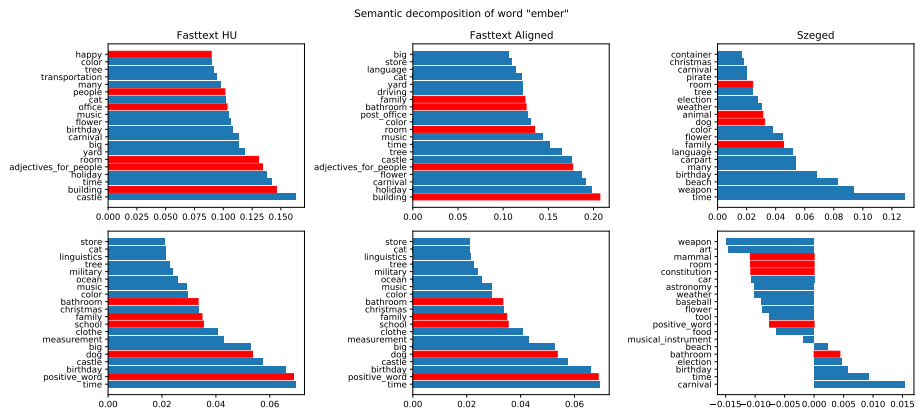


Fig. 4. Semantic decomposition of the word "ember". First row shows the decomposition of dense embedding spaces and the second represents the sparse embedding spaces ($k = 2000$, $\lambda = 0.05$). On the y axis we represent the semantic categories and on the x axis we show the corresponding weights of the word. Red bars represents that if the word is in the semantic category.

the interpretability of the Hungarian Fasttext, Hungarian Aligned Fasttext, and Szeged WV models as source embeddings, where we concluded that all of them are capable to express the anticipated semantic information contents and that the aligned word vectors performed above all. Furthermore, we proposed a modified version of the interpretability score, which let us compare the interpretability of embedding spaces with different dimensionality and consider errors from the transformation.

We also considered the utilization of the Hellinger distance instead of Bhattacharyya distance which improved the interpretability scores. Furthermore, we explored the behavior of sparse representations. As for the hyperparameter selection, we can conclude that we want to increase the number of the basis, and decrease the sparsity level in order to improve the performance.

However, if we consider sparse representations the generalisability of the embedding may decrease, but it might be a joint factor of the distant supervised generation of Hungarian semantic categories and random selection of validation test sets. If our semantic categories contain too much noise then it could accumulate that noise during the transformation which is indicated by the high interpretability score, and a lower score on the word retrieval test (which can represent a distinct distribution from the original distribution of the semantic category).

Acknowledgments

The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program and the Artificial Intelligence National Excellence Program (grant no.: 2018-1.2.1-NKP-2018-00008). The work of Tamás Ficsor was funded by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002, supported by the EU and co-funded by the European Social Fund.

Bibliography

- Abka, A.: Evaluating the use of word embeddings for part-of-speech tagging in bahasa indonesia. pp. 209–214 (10 2016)
- Balogh, V., Berend, G., Diochnos, D.I., Turán, Gy.: Understanding the semantic content of sparse word embeddings using a commonsense knowledge base (2019)
- Berend, G.: Sparse coding of neural word embeddings for multilingual sequence labeling. *Transactions of the Association for Computational Linguistics* 5, 247–261 (2017), <https://www.aclweb.org/anthology/Q17-1018>
- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5, 135–146 (2017)
- Boleda, G.: Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6(1), 213–234 (Jan 2020), <http://dx.doi.org/10.1146/annurev-linguistics-011619-030303>
- Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017)
- Forgues, G., Pineau, J., Larchevêque, J.M., Tremblay, R.: Bootstrapping dialog systems with word embeddings. In: *Nips, modern machine learning and natural language processing workshop*. vol. 2 (2014)
- Friedman, J., Hastie, T., Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)* 9, 432–41 (08 2008)
- Ganchev, K., Graça, J.a., Gillenwater, J., Taskar, B.: Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.* 11, 2001–2049 (Aug 2010)
- Garrard, P., Ralph, M., Patterson, K.: Prototypicality, distinctiveness, and inter-correlation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive neuropsychology* 18, 125–74 (03 2001)
- Harris, Z.S.: Distributional structure. *WORD* 10(2-3), 146–162 (1954), <https://doi.org/10.1080/00437956.1954.11659520>
- Joulin, A., Bojanowski, P., Mikolov, T., Jégou, H., Grave, E.: Loss in translation: Learning bilingual word mapping with a retrieval criterion. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018)
- Kazama, J., Tsujii, J.: Evaluation and extension of maximum entropy models with inequality constraints pp. 137–144 (01 2003)

- Kenter, T., de Rijke, M.: Short text similarity with word embeddings. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. pp. 1411–1420. CIKM '15, ACM, New York, NY, USA (2015), <http://doi.acm.org/10.1145/2806416.2806475>
- Lebret, R., Collobert, R.: Word embeddings through hellinger pca. Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (2014)
- Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. vol. 382, p. 87 (01 2009)
- Martins, A., Smith, N., Figueiredo, M., Aguiar, P.: Structured sparsity in structured prediction. pp. 1500–1511 (01 2011)
- McRae, K., Cree, G., Seidenberg, M., McNorgan, C.: Semantic feature production norms for a large set of living and nonliving things. Behavior research methods 37, 547–59 (12 2005)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality (2013)
- Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011. Association for Computational Linguistics, Suntec, Singapore (Aug 2009), <https://www.aclweb.org/anthology/P09-1113>
- Mohd, M., Jan, R., Shah, M.: Text document summarization using word embedding. Expert Systems with Applications 143, 112958 (2020), <http://www.sciencedirect.com/science/article/pii/S0957417419306761>
- Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116(44), 22071–22080 (Oct 2019), <http://dx.doi.org/10.1073/pnas.1900654116>
- Murphy, B., Talukdar, P., Mitchell, T.: Learning effective and interpretable semantic models using non-negative sparse embedding. In: Proceedings of COLING 2012. pp. 1933–1950. The COLING 2012 Organizing Committee, Mumbai, India (Dec 2012), <https://www.aclweb.org/anthology/C12-1118>
- Nugaliyadde, A., Wong, K.W., Sohel, F., Xie, H.: Enhancing semantic word representations by embedding deeper word relationships. CoRR abs/1901.07176 (2019), <http://dblp.uni-trier.de/db/journals/corr/corr1901.html>
- Park, S., Bak, J., Oh, A.: Rotated word vector representations and their interpretability. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 401–411. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017), <https://www.aclweb.org/anthology/D17-1041>
- Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation pp. 1532–1543 (Oct 2014a)
- Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014b)

- Seok, M., Song, H.J., Park, C.Y., Kim, J.D., Kim, Y.S.: Named entity recognition using word embedding as a feature. *International Journal of Software Engineering and its Applications* 10, 93–104 (2016)
- Shen, Y., Rong, W., Nan, J., Peng, B., Tang, J., Xiong, Z.: Word embedding based correlation model for question/answer matching (11 2015)
- Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: *AAAI Conference on Artificial Intelligence* (2016), <http://arxiv.org/abs/1612.03975>
- Subramanian, A., Pruthi, D., Jhamtani, H., Berg-Kirkpatrick, T., Hovy, E.: Spine: Sparse interpretable neural embeddings (2017)
- Szántó, Z., Vincze, V., Farkas, R.: Magyar nyelvű szó-és karakterszintű szóbeágyazások. XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2018), Szeged, Szegedi Tudományegyetem, Szegedi Tudományegyetem pp. 323–328 (2017)
- Turian, J., Ratinov, L.A., Bengio, Y.: Word representations: A simple and general method for semi-supervised learning. In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. pp. 384–394 (2010)
- Véronis, J.: Hyperlex: lexical cartography for information retrieval. *Comput. Speech Lang.* 18(3), 223–252 (2004), <http://dblp.uni-trier.de/db/journals/csl/csl18.html#Veronis04>
- Yogatama, D., Smith, N.A.: Linguistic structured sparsity in text categorization. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 786–796. Association for Computational Linguistics, Baltimore, Maryland (Jun 2014), <https://www.aclweb.org/anthology/P14-1074>
- Zou, W.Y., Socher, R., Cer, D., Manning, C.D.: Bilingual word embeddings for phrase-based machine translation. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pp. 1393–1398. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013), <https://www.aclweb.org/anthology/D13-1141>
- Şenel, L.K., Utlü, I., Yücesoy, V., Koç, A., Çukur, T.: Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26(10), 1769–1779 (2018)
- Şenel, L.K., Utlü, I., Şahinuç, F., Ozaktas, H.M., Koç, A.: Imparting interpretability to word embeddings while preserving semantic structure. *Natural Language Engineering* p. 1–26 (2020)

Interaktív tematikus-szemantikus térkép a Történeti Magánéleti Korpusz keresőfelületén

Novák Attila^{1,2}

¹Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

²MTA-PPKE Magyar Nyelvtudományi Kutatócsoport

Budapest, Práter u. 50/a.

{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat A cikkben a TMK Történeti Magánéleti Korpusz webes lekérdezőfelületének újdonságait mutatjuk be, különös tekintettel a korpusz lexikai anyagát szemléltető, szóbeágyazási modellek felhasználásával készített interaktív tematikus-szemantikus térképekre. A pusztán a TMK-ból készített, a korpusz kis mérete miatt jellegében inkább tematikusnak, mint igazán nyelvinek mondható szóbeágyazási modell mellett a TMK kibővítésével nyert korpuszból készített már inkább nyelvi-szemantikus modellekből a t-SNE algoritmussal nyert kétdimenziós lexikai térképek elemeire kattintva közvetlenül is indítható az adott nyelvi elemre vonatkozó korpuszlekérdezés. A térképek ugyanakkor a szövegek gépi feldolgozásakor, illetve kézi ellenőrzésekor bent maradt hibákra is felhívják a figyelmet, könnyítve ezzel a hibajavítást.

Kulcsszavak: interaktív vizualizáció, t-SNE, szóbeágyazási modellek, korpuszlekérdező, történeti korpusz

1. A Történeti Magánéleti Korpusz

A Történeti Magánéleti Korpusz (TMK)¹ két OTKA, illetve NKFIH kutatási pályázat² keretében jött létre a Nyelvtudományi Intézetben. A második pályázat 2020-ban ért véget. A TMK ó- és középmagyar korból származó olyan írott nyelvemlékekből áll, amelyek a magánéleti nyelvi regiszterhez legközelebb állónak tekinthetők. 1772 előtt keletkezett magánlevelek és perjegyzőkönyvek képezik a korpusz anyagát nagyjából azonos arányban. Elsősorban történeti morfológiai, szociolingvisztikai, történeti mondattani, pragmatikai és lexikológiai vizsgálatokat szem előtt tartva állítottuk össze a korpusz anyagát, és ezek a szempontok határozták meg az annotációs elveket is. A korpusz mérete a második pályázat

¹ <http://tmk.nytud.hu/>

² OTKA K 81189: *Morfológiailag elemzett nyelvtörténeti korpusz a magánéleti nyelvhasználat köréből* (2010-2014), NKFI-OTKA K 116217: *Versengő szerkezetek a középmagyar élnyelvben: változók elemzésén alapuló megközelítés* (2015-2020). Mindkét kutatás vezetője Dömötör Adrienne volt.

zárultakor 8,6 millió karakter, ebből 7,7 millió karakter magyar nyelvű. A magyar nyelvű rész teljes egészében morfológiailag annotált, ez összesen 1 millió 112 ezer elemzett szövegszó.³

A korpusz nyomtatott forráskiadások feldolgozásával készült.⁴ A szövegek eredeti alakját a nyomtatott kiadásokban szereplő formában vettük át. Ez a szövegváltozat OCR-ezéssel és az így digitalizált szöveg kézi javításával állt elő. A szövegeket félautomatikus módon tagmondatokra bontottuk, majd a tagmondatokra bontást kézzel javítottuk. A más tagmondatokba beágyazott tagmondatokat külön megjelöltük. A tagmondatokra bontott szövegekhez kézzel a mai magyar helyesírásra normalizált változat készült. A normalizálás során neutralizáltuk a morfológiai következményekkel nem járó tisztán fonológiai jellegű nyelvjárási sajátosságokat, de nem változtattunk a szavak morfológiai szerkezetén: a történeti szövegekre jellemző morfológiai és szintaktikai szerkezeteket a normalizálás nem érintette.

A szövegeket a Humor morfológiai elemző (Novák, 2003) ó- és középmagyar szövegekre adaptált változatával (Novák és Wenszky, 2013) elemeztük morfológiailag, és a PurePos szófaji egyértelműsítő eszközzel (Orosz és Novák, 2013) egyértelműsítettük automatikusan. A géppel elemzett és egyértelműsített szövegeket egy erre a célra készült webes egyértelműsítő felületen manuálisan ellenőriztük és javítottuk. Itt az esetleges elemzési illetve egyértelműsítési hibákon kívül a normalizálási, tokenizálási és tagmondatokra bontási hibákat is javítani lehet, és a javításuk után a javított részeket újra lehet elemezteni. A kézi ellenőrzésen és javításon átesett szövegek a projekt előrehaladása folyamán folyamatosan bekerültek a PurePos egyértelműsítő tanítóanyagába. A Humor morfológiai elemző lexikonját is folyamatosan bővítettük az újonnan elkészült normalizált szövegek szóanyagával. A projekt folyamán nem került sor a korpusz teljes elemzett anyagának kézi ellenőrzésére: az elemzések 78%-a van kézzel ellenőrizve.⁵

A szociolingvisztikai szempontú kutatások segítése érdekében minden szöveget annotáltunk a rendelkezésre álló metaadatokkal. Ezek között minden esetben megtalálható az adott szöveg keletkezésének éve, illetve a levelek esetében pontos dátuma, a keletkezés helye, perek esetében a megye és a műfaj (levél, illetve per). A levelek esetében emellett a szerző, illetve a címzett neve, neme, illetve társadalmi státusza, a szerző és a címzett közötti viszony jellege, valamint az adott szövegrész saját kezű mivoltára vonatkozó információ szerepel a metaadatok között. Ezen kívül az egyes szövegrészeket annotáltuk a szövegrész típusa szerint a szövegtörzs mellett megkülönböztetve a címzést, a külső, a margón tett megjegyzéseket és a mellékleteket, illetve perek esetében a formulaszerű hivatalos részeket.

³ Ez valóban ennyi szót és nem token-t jelent, az írásjeleket nem tekintettük külön tokennek.

⁴ <http://tmk.nytud.hu/forrasok.php>

⁵ Korábbi méréseink során (Dömötör és mtsai, 2017) a gépi egyértelműsítés pontossága a szótokenek szintjén az írásjeleket nem figyelembe véve 95,9%-osnak, a tagmondatok szintjén 81,5%-osnak adódott (a tagmondatok ötödében találtunk hibát).

Alapesetben a találati egységek mondatok, amelyek tagmondatokra vannak bontva és a találatot adó szavak ki vannak emelve. Ez a kiemelés az Emdros terminológiájában a *fókusz*: a példában a *bízik* alakjai. Alapesetben a mondatok interlineáris formátumban jelennek meg (1. és 2a ábra) és a szavak eredeti és normalizált alakját, szótövét és a morfoszintaktikai annotációját külön sorokban tartalmazzák. A megszakított tagmondatokat eltérő háttérszín jelzi. Minden egyes mondattalálát fölött szerepelnek a találatot adó szöveg főbb jellemzői. A szövegazonosító mellett a dátum, szerző és címzett, illetve a per helyszíne, a szerző és a címzett viszonya (az 1. ábrán levelekből, a 2. ábrán perszövegből származó találatokat látunk). Itt jelezzük emellett, hogy az adott szöveg átesett-e a gépi annotációt követő kézi ellenőrzésen (E=ellenőrzött, NE=nem ellenőrzött). A találathoz tartozó metaadatokra kattintva külön ablakban a teljes szöveg megnyílik, amelyen belül a keresésben találatot adó szavak ugyanúgy ki vannak emelve, mint az eredeti egymondatos találatokban. A teljes annotáció mellett a találatok egyszerűsített formában morfológiai annotáció nélkül is megjeleníthetők. Ebben a változatban választható, hogy a találatokat az eredeti (2b ábra) vagy a normalizált alakjukban szeretnénk látni (2c ábra). A teljes mondatos találatok mellett gyakorisági adatok is kérhetők a rendszertől. Ilyenkor megadható, hogy a találati elemeknek melyik jellemzői jelenjenek meg.

[1] Bosz. 1a. Abaúj-Torna megye, Szilas, 1736. ::: (E) - 1063682															
125173	125174	125175	125176	125177	125178	125179	125180	125181	125182	125183	125184	125185	125186	125187	125188
egy	kis	idő	múlva	estve fell	még	világos	völt	Tehin gyüvéskor	gyön	Falubul	edgy	nagy	Files Bagoly	nagy	csetajjal patajval,
Egy	kis	idő	múlva,	estefelá,	<még	világos	völt.>	tehnjövés	jön	faluból	egy	nagy	fülesbagoly	nagy	csetajjal-patajjal,
Det	Adj	N	PP	Adv	Adv	Adj	V.Past.S3	N.Tem	V.S3	N.Ela	Det	Adj	N	Adj	N.Ins
125189	125190	125192			125193	125194	125195	125196	125197	125198	125199	125200			
fel	az	úton	mentiben			ahol	a	szőlő	közt	völt.	oda gyött	igenessen	hozzája,		
fel	az	úton	mentiben,			<ahol	a	szőlő	közt	völt.>	odajött	egyenesen	hozzája,		
VPtx	Det	N.Sup	megy			a+hol	a	szőlő	közt	van	oda+jön	egyenes	ő		
			V_Nact=IAPxS3.Ine			Adv Pro Rel	Det	N	PP	V.Past.S3	VPtx.V.Past.S3	Adj.Essmod	N Pro.All.S3		

(a) Interlineáris megjelenítés - beágyazott tagmondatokkal

[1] Bosz. 1a. Abaúj-Torna megye, Szilas, 1736. ::: (E) :::																				
egy	kis	idő	múlva	estve feli	.	még	világos	völt	.	Tehin	gyüvéskor	gyön	Falubul	edgy	nagy	Files	Bagoly	nagy	csetajjal patajval,	.
fel	az	úton	[mentiben]	.	ahol	a	szőlő	közt	völt.	.	oda	gyött	igenessen	hozzája,	.					.

(b) Egyszerűsített megjelenítés - eredeti alak

[1] Bosz. 1a. Abaúj-Torna megye, Szilas, 1736. ... (E)

Egy kis idő múlva, estefelé, . <még világos volt.> . tehénjövéskor jön faluból egy nagy fülesbagoly nagy csetajjal-patajjal, . fel az úton [mentében,] . <ahol a szőlő között volt.> . odajött egyenesen hozzája.

(c) Egyszerűsített megjelenítés - normalizált alak

2. ábra. Megjelenítési formátumok a korpuszlekérdezőben

A kereső speciális szolgáltatása, hogy a megfelelő jogosultsággal rendelkező felhasználók számára lehetővé teszi a keresőben való hibajavítást is (3. ábra). Egy adott szóra kattintva a kézi egyértelműsítő felülethez hasonló módon javítható a szó eredeti, illetve normalizált alakja, elérhető a morfológiai elemző, melynek elemzéseit közül választhatunk, illetve kézzel is szerkeszthetjük az elemzés. Emellett a tokenizálási és tagmondatokra bontási hibák javítására is van lehetőség.

31

1053560	1053561	1053562	1053563	1053564	1053565
Nador Jspannyanak	etc	ennekem	Zerette	Bízotth	Vramnak
nádorispánjának	etc.,	énnekem	szerette	bízott	uramnak.
nádor+ispán	etc.	én	szeret[V.PartPrf_Subj=tA.PxS3]	bízott	úr
N.PxS3.Dat	Inj/Utt	N Pro.Dat.S1	♥ OK < > X <>	Adj	N.PxS1.Dat

52	667353	667354	667355	667356	667357	667358	667359	667360	667361	667362
	az	dologh,	reménlem	továb	is	vigasztalássára	leszen	az	egész	hazának.
	a	dolog,	reménylem,	tovább	is	vigasztalására	leszen	az	egész	hazának.
	a	dolog	reményli[V.S1.Def]	tovább	is	vigasztalás	lesz	az	egész	haza
st.S3	Det	N	♥ OK < > X <>	Adv	Clit_is	N.PxS3.Sub	V.S3	Det	Adj	N.Dat

5	386596	386597	386598	386599	386600	386601	386602
	talán	az	cselekedte	az	leányán	történt	nyavalát.
	talán	az	cselekedte	a	leányán	történt	nyavalát.
	talán	az	cselekszik	a	leány	történik[V.Past.S3]	nyavalatya
Adv	N Pro		V.Past.S3.Def	Det	N.PxS3.Sup	♥ OK	c
						történik[V.Past.S3]	
						történik[V.PartPrf]	

494964	494965	494966	494967	494968	494969	494970	494971	494972	494973
--------	--------	--------	--------	--------	--------	--------	--------	--------	--------

3. ábra. Hibajavítás a TMK kereső találatában.

3. Szemantikus térképek a korpusz lexémáiról

A projektum zárószakaszában a keresőt egy új szolgáltatással egészítettük ki. Ez neurális disztribúciós modelleken alapuló kétdimenziós lexikális térképekből áll, amelyek a korpuszban legalább háromszor előforduló lexikai elemek disztribúciós szemantikai térben való reprezentációját vizualizálják.⁷ A egyes elemek a gyakoriságukkal (logaritmikusan) arányos méretben és a szófajukra jellemző színben jelennek meg (5., 7., 8. ábrák). A térképböngészőbe keresési funkciót is integráltunk, amelynek segítségével a térképen szereplő lexikai egységekre kereshetünk illeszkedő szórészletek alapján. A találatok átmenetileg kiemelt színnel és kinagyítva jelennek meg, illetve egyenként végiglépegethetünk rajtuk az adott elem környékére automatikusan ráközelítve. A térképek annyiban interaktívak, hogy a rajtuk szereplő lexikai elemekre duplán kattintva lekérdezés kezdeményezhető az adott elemre a korpuszból. A lekérdezés eredménye új böngészőfülön jelenik meg (9. ábra).

3.1. Előzmények

Korábban milliárdszavas nagyságrendű magyar nyelvű webkorpuszból hoztunk létre a word2vec (Mikolov és mtsai, 2013), illetve a fastText (Bojanowski és mtsai, 2016) eszköz CBOW modelljével háromszáz dimenziós disztribúciós modelleket. Nyers szövegen tanított modellek mellett morfológiailag annotált szövegen be-tanított modelleket is létrehoztunk, amelyek a ritkább szavakra jobb minőségű reprezentációt hoztak létre, mert a lemmatizálás csökkentette az adatritkaságot

⁷ <http://tmk.nytud.hu/maps.php>

(Novák és Novák, 2018). Azokban a modelljeinkben, amelyekben a fő szófajcím-két is a lemmatizált elemek részévé tettük, a módszer azon hiányosságát is sikerült részben kiküszöbölni, hogy önmagában nem alkalmas a homonímia, illetve poliszmia kezelésére.⁸ A modelljeinket korábban t-SNE (t-distributed stochastic neighbor embedding) algoritmus (van der Maaten és Hinton, 2008) segítségével vizualizáltuk és jelen kutatásban is ezt a módszert alkalmaztuk.

A korábban létrehozott sok millió lexikai elemet tartalmazó modelljeink esetében a vizualizációt a modellt böngésző felhasználó által menet közben összeállított korlátos szókészletre dinamikusan hoztuk létre a szerveren (Novák és mtsai, 2017). Mivel a t-SNE algoritmus gradiens ereszkedés algoritmussal (SGD) optimalizálja a képet eloszlások Kullback–Leibler (KL)-távolságát hibafüggvényként használva,⁹ ezért futtatása a szerveren meglehetősen idő- és erőforrás-igényes (sok ezer pont megjelenítése esetén a keresést futtató szerveren több percre tartat az ábra generálása). Ezt a TMK keresőfelületére integrálandó interaktív vizualizáció esetében mindenképp szeretnénk volna elkerülni. Korábban kísérleteztünk autoenkoderen alapuló vizualizációval is, amely a képgenerálás válaszüdejét jelentősen csökkenthetné, ez azonban a szóbeágyazási modellen alapuló szemantikus térképek megjelenítésére nem adott elfogadható minőségű megoldást (Novák és Novák, 2020).

3.2. A TMK felületén alkalmazott megoldás

Ugyanakkor a szemantikus térkép megjelenítése a kliens gép böngészőjében szintén túlzott erőforrásigényt jelent, ha a modell túl nagy. Ez a modellt nézegető felhasználó gépén a böngésző vagy akár a teljes operációs rendszer reszponzivitásának megszűnéséhez vezethet a túlzott memóriaigény miatt. Ezért olyan megoldást kellett találni, amely sem a szerveret, sem a kliensgépet nem terheli túl. Ezt úgy tudtuk megoldani, hogy a kétdimenziós térképeket offline legeneráltuk, de a modell méretét úgy korlátoztuk, hogy az ábra megjelenítése és böngészése legalábbis egy nem túl korlátozott memóriakapacitású klienskonfiguráció esetén ne jelentsen gondot.¹⁰ A megjelenítendő modellt a korpuszban legalább háromszor előforduló szófajkóddal annotált lemmák képére korlátozva elfogadható modellméretet kaptunk (13500 lexikai elem). A szemantikus térképek megjelenítését végző kódot, amely a javascript-alapú cytoscape.js gráfvizualizációs és -szerkesztő csomagon alapul (Franz és mtsai, 2015), a Novák és Novák (2020)-ban bemutatott kód adaptálásával készítettük el.

⁸ Az esetleges elemzési hibáktól eltekintve ennél a korpuszméretnél a különböző szófajú lemmák szétválasztása egyértelműen jelentős mértékben javítja a modell minőségét, és nem vezet adatritkasági problémákhoz.

⁹ Az eredeti modellbeli távolságokkal arányos feltételes valószínűségeket adó gaussi eloszlások és a párdimenziós kép pontjai közötti távolságokat adó Student t-(Cauchy)-eloszlások közötti KL-távolságot optimalizálja. Erre utal módszer nevében a *t*.

¹⁰ 4GB RAM-mal szerelt laptopon Chrome böngészőben problémamentesen működik.

4. A modellek előállítása

4.1. Az elemzett TMK-n betanított modell

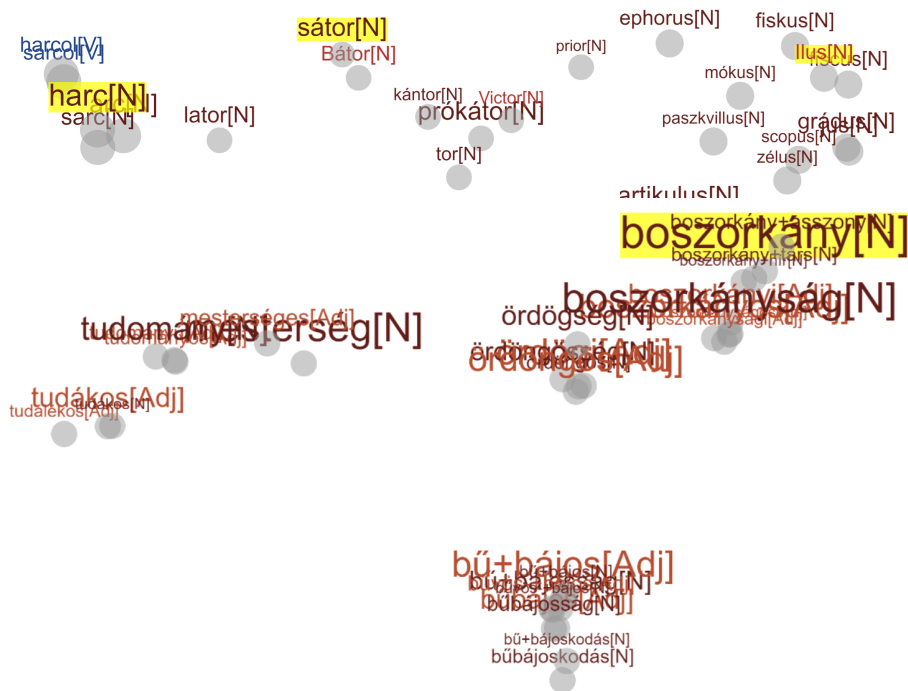
A TMK korpusz a korábbi kísérleteinkhez használt korpusznál három nagyságrenddel kisebb méretű, ezért a korábban alkalmazott módszerek még a lemmatizálással együtt sem adtak a nagy korpuszon kapott modellhez hasonló minőségű eredményt. Valamelyest enyhített a problémán, hogy a modellek létrehozására a fastText eszközt használtuk, amely nem tokenek, hanem karakter-n-gramok reprezentációját hozza létre, így a tanítóanyagban nem szereplő szavakhoz is létre tud hozni reprezentációt a szót alkotó n-gramok reprezentációjának átlagolásával. Emellett kevés minta esetén is viszonylag értelmes eredmény jöhet ki, ha a hasonló szavaknak valóban van közük egymáshoz. Ugyanakkor ez a megközelítés kevés minta esetén reprezentációs problémákhoz vezethet a véletlen hasonlóságok esetében. Pusztán a kb. egymillió szavas elemzett TMK korpuszon betanítva a modell nem volt képes arra, hogy a ritka szóalakokhoz a néhány előfordulásuk alapján megfelelő reprezentációt hozzon létre, ezért az ilyen elemekhez a legközelebbi szomszédok lekérdezésekor leginkább a hasonló karaktersorozatokat tartalmazó, de nyelvi nem feltétlenül releváns találatok jönnek ki. Gyakoribb szavaknál is sokszor inkább a tematikus, mint a nyelvi hasonlóságok dominálnak (4. és 5. ábra).

0	harc[N]	1	30	0	sátor[N]	1	36	0	öreg[Adj]	1	282	0	apa[N]	1	117
1	harc[N]	1.00000005607	30	1	sátor[N]	1.0000	36	1	öreg[Adj]	1.0000	282	1	apa[N]	1.000000002151	117
2	arc[N]	0.8127	15	2	Bátor[N]	0.7109	16	2	öreg[N]	0.8150	10	2	papa[N]	0.7644	5
3	sarc[N]	0.7738	6	3	szenátor[N]	0.6908	4	3	öregség[N]	0.6287	6	3	apá[N]	0.7428	3
4	harc+hely[N]	0.7506	3	4	tor[N]	0.6446	9	4	öreg+leány[N]	0.5572	9	4	kapa[N]	0.7406	16
5	bérc[N]	0.7011	3	5	tutor[N]	0.6301	5	5	öreg-bíró[N]	0.5520	5	5	a[N]	0.7375	9
6	harcol[V]	0.6879	5	6	sátán[N]	0.6213	5	6	agg[Adj]	0.5465	3	6	nagy+apa[N]	0.7357	3
7	hab[N]	0.6382	6	7	sás[N]	0.6201	3	7	öreg+ember[N]	0.5417	7	7	kupa[N]	0.7251	3

4. ábra. Néhány legközelebbi szomszéd a pusztán a TMK-ból generált modellben.

4.2. Módosított algoritmus

A problémákat úgy próbáltuk orvosolni, hogy további tanítóanyaggal egészítettük ki a korpuszt. Itt azonban problémát jelentett, hogy a hozzáadott tanítóanyagot is a korpusz elemzésével kompatibilis elemzéssel kellett ellátni ahhoz, hogy annotált anyagon alapuló modellt tudjunk létrehozni. Felmerült az az ötlet, hogy az algoritmus módosításával esetleg elemzetlen szöveget is lehetne használni. Ehhez a kísérlethez a fastText CBOW algoritmusának módosított változatát használtuk (CBOW/A), amely alkalmas olyan vektortérmodell létrehozására, amely egyszerre tartalmazza a felszíni szóalakok és az elemzett lemmák reprezentációját (Novák és mtsai, 2019). Az algoritmus alkalmazásához olyan korpuszreprezentációra van szükség, amely a felszíni alakok mellett azok valamilyen annotált



változatát is tartalmazza (1c). Az annotációkat konfigurálható prefix jelöli (a példában: .). A pusztán az elemzéseket tartalmazó modell készítéséhez az eredeti CBOW algoritmus használatakor korábban a (1b)-ben látható formátumot használtuk a modell tanításához.

- (1) a. Szeretettel való szolgálatomat ajánlom kegyelmednek, édes szívem!
 b. szeretet[N] [Ins] való[Adj] szolgálat[N] [PxS1.Acc] ajánl[V] [S1.Def] kegyelme[N|Pro] [PxS2.Dat] , édes[Adj] szív[N] [PxS1] !
 c. Szeretettel .szeretet[N] való .való[Adj] szolgálatomat .szolgálat[N] ajánlom .ajánl[V] kegyelmednek .kegyelme[N|Pro] , édes .édes[Adj] szívem .szív[N] !

Tanításkor az algoritmus véletlenszerűen mintavételezi az egyes korpuszpozíciókban a felszíni szóalakot és az adott pozícióhoz tartozó annotációt, így a tanítás során a korpuszon többször végigmenve a felszíni szóalakok és az annotációk reprezentációja is létrejön. A mi konkrét esetünkben a felszíni szóalakok a normalizált alakok, az annotációt pedig a szófajcímkével ellátott lemmák alkotják.

A CBOW/A algoritmust pusztán a TMK-n futtatva a lemmák modellbeli képe nem javult észrevehető módon, így önmagában az algoritmus lecserélése nem

javított a modell minőségén. Abban reménykedtünk azonban, hogy a tanítókorpusz bővítésével az n -gramok jobb reprezentációja segítheti a TMK lemmáinak jobb minőségű ábrázolását is.

4.3. A korpusz bővítése

A korpuszt olyan mai magyar szövegekből vett mondatokkal egészítettük ki, amely olyan szavakat tartalmaz, amelyek a TMK korpuszban is előfordultak, de 100-nál kevesebb előfordulásuk volt. Az új tanítóanyag első változata úgy állt össze, hogy a webkorpuszból szűrt anyagot elemzetlenül adtuk hozzá a TMK elemzett annotált anyagához. A webkorpuszból vett kiegészítés mérete 41,8 millió szó volt. A modell felépítése után azt visszaszűrtük csak a TMK szavaira.

Ebben a modellváltozatban a felszíni szóalakok legközelebbi szomszédait megnézve azt láttuk, hogy azok reprezentációja valóban nagyon sokat javult, mert a TMK-ban ritkább szóalakokra sok példa volt a bővített korpuszban. Azonban visszaszűrve a lemmák reprezentációjára semmilyen lényegi javulást nem láttunk ezek minőségében. Ráadásul a szóalakok reprezentációja nagyon eltávolodott a lemmákétól.

Ezért a következő modellváltozat elkészítéséhez a webes korpuszból vett anyagot is leelemztük a TMK elemzéséhez használt elemzőlánccal. Ezután az így kapott modellt is visszaszűrtük csak a TMK szavaira. Ebben a modellben a lemmák legközelebbi szomszédait megnézve azt találtuk, hogy a lemmák reprezentációja is elfogadható minőségűre javult (6. és 7. ábra).

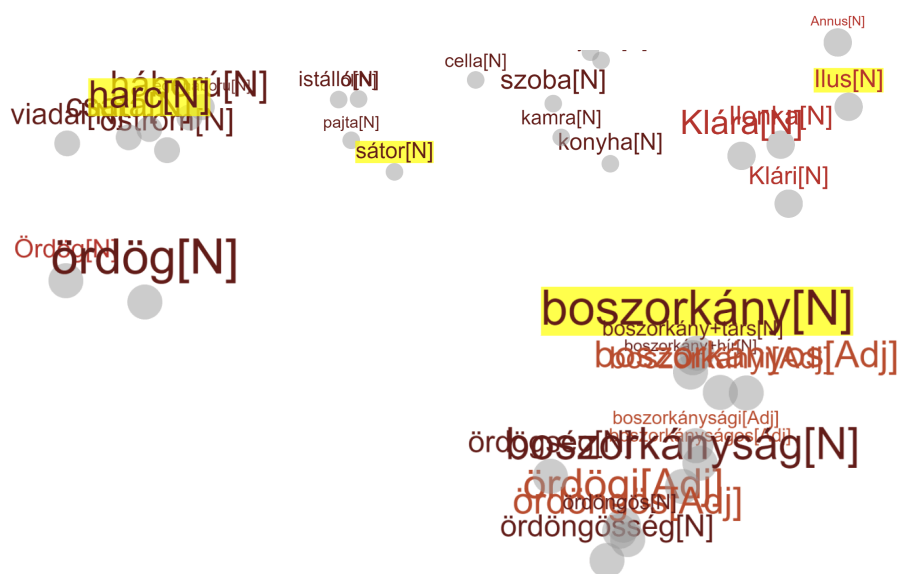
0	harc[N]	1	2851	0	sátor[N]	1	895	0	öreg[Adj]	1	3024	0	apa[N]	1	5353
1	harc[N]	1.00000006884	2851	1	sátor[N]	1.00000011473	895	1	öreg[Adj]	1.00000006603	3024	1	apa[N]	1.0000	5353
2	csata[N]	0.7949	1661	2	szoba[N]	0.6728	4290	2	öreg[N]	0.8053	579	2	anya[N]	0.8621	8722
3	háború[N]	0.7603	2624	3	ház[N]	0.6591	16842	3	vér[Adj]	0.7576	693	3	édes+apa[N]	0.8062	1159
4	harc+hely[N]	0.7445	10	4	asztal[N]	0.6455	4927	4	öreg+ember[N]	0.7264	238	4	édes+anya[N]	0.7924	2425
5	elenség[N]	0.7207	2500	5	láda[N]	0.6424	908	5	idő[Adj]	0.7261	2911	5	nagy+apa[N]	0.7924	644
6	támadás[N]	0.7203	1754	6	kocsi[N]	0.6417	4385	6	szegény[Adj]	0.6960	4075	6	nagy+anya[N]	0.7919	407

6. ábra. Néhány legközelebbi szomszéd a kiegészített korpuszból generált modellben.

Végeztünk egy harmadik kísérletet is, amelyben azt vizsgáltuk meg, hogy van-e jelentősége, hogy menet közben a modellt a TMK-ban nem szereplő lemmákra is betanítjuk. Ebben a kísérletben a tanítóanyag úgy állt elő, hogy a kiegészítő korpuszban a TMK-ban nem szereplő szavak elemzését eldobtuk és így tanítottuk be a neurális modellt. Majd a modellt itt is visszaszűrtük csak a TMK szavaira. A modellbe tekintve azt találtuk, hogy ez a megközelítés az előző változathoz hasonló modellt eredményezett.

4.4. A lexikai térképek

Az így előállt modellekből a kétdimenziós térképeket előállítva azt találtuk, hogy tulajdonképpen a pusztán a TMK-ból a korpusz (1b)-ben látható elemzett alak-

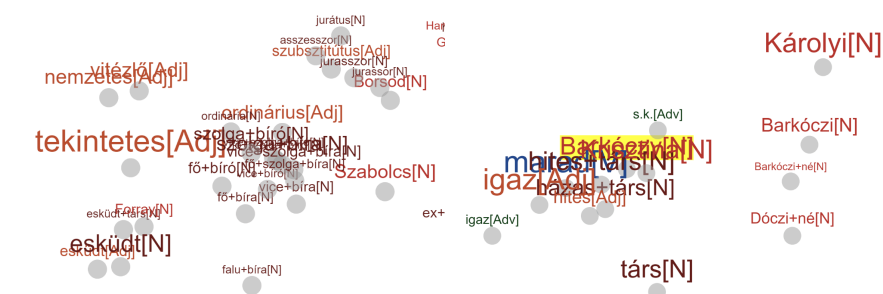


7. ábra. Néhány részlet a bővített korpuszból generált modellben.

jából az eredeti CBOW algoritmussal készített lemma modell is jól használható áttekintést ad a korpusz szavairól, amelyben azonban nem elsősorban a nyelvi jellegű szerveződési szintek alapján csoportosulnak az elemek, hanem sok helyen inkább a korpuszra jellemző tematikus csoportok dominálnak. A korpusz anyagát ismerő kutatók számára hamar feltűnnek azok a sűrűsödési pontok a térképen, ahol az egyes részkorpuszok nagyon jellemző fordulatai, nevei csoportosulnak (8. ábra). Emellett helyenként pusztán a sztringhasonlóság hozza egymáshoz közel az elemeket, amely ugyan az esetek nagy részében az egymáshoz közel lévő elemek nyelvi hasonlóságával jár együtt, de időnként egyszerűen csak rímelnek az egymáshoz közel lévő szavak hasonlóan az 5. ábrán felül látható esetekhez (pl. *arc-harc-sarc*).

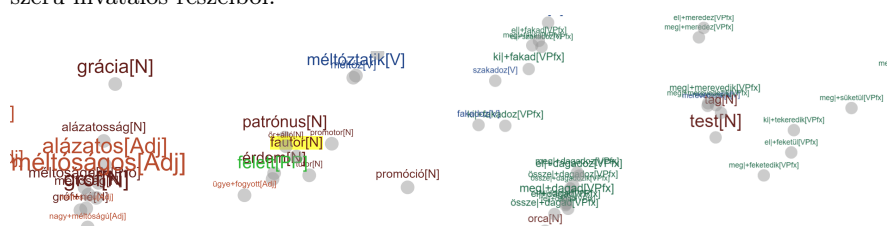
A kiegészített modellekből készült térképeken sokkal inkább a lexikai tér nyelvi szerveződése érhető tetten. A különböző szófajú elemek nem keverednek olyan mértékben egymással, mint a kis modellből készített térképen. Helyenként tetten érhető, hogy a mai korpuszban domináló jelentések irányába mozdult el a kép (pl. míg a pusztán a TMK-ból készített térképen a *mesterséges* és a *tudományos* boszorkánysághoz köthető fogalmakként jelennek meg, a *közösködik* a *közösül* szinonimája, a kiegészített korpuszból készült modellből generált térképen ezeknek a lexémáknak a képe elmozdult a mai jelentésüknek megfelelő helyekre.)

A térképeken időnként meglepő helyeken jelennek meg meglepő lexikai elemek. Ezeknek a jelenségeknek könnyen utána járhatunk az adott elemre kattintva kapott lekérdezések eredményére rátekintve, és azt találjuk, hogy az elemzett korpusz kézzel nem ellenőrzött részéből származó reprezentációról van szó (9. ábra). Például a női nevek között feltűnő *bután* főnév az egyik boszorkányper-



(a) A jogászszereplőket leíró jellegzetes fordulatok szavai a perszövegek formula-szerű hivatalos részeiből.

(b) A Károlyi-Barkóczi-levelezés jellegzetes fordulatainak elemei.



(c) A peregrinuslevelek jellegzetes szavai.

(d) Röntások hatásai a boszorkányperek-ből, és amit érint.



(e) A boszorkányperek jellegzetes szavai.

8. ábra. Néhány tematikus sűrűsödési pont a TMK-ból az eredeti CBOW algoritmussal generált modellben.

szereplő, Butáné ‘vices’ elemzéséből adódik, hasonlóan a *lakol* közelsége a *lakik*-hoz jól mutatja, hogy valójában nem a *(meg)lakol* igéről van szó, hanem a lakik szubsztenderd második személyű alakjáról, amely néhány ellenőrizetlen szövegben hibás elemzéssel maradt benne. A modell azokat az eseteket is felszínre hozza, ahol a normalizálás során a fonológiai variabilitásból adódó különbségeket nem sikerült teljesen semlegesíteni (pl. *csujtár-csótár* nyereg alatti lótakaró’).



9. ábra. Néhány példa elemzési/normalizálási hibákra. Az elem helye a térképen utal arra, hogy hogyan kell javítani.

5. Összefoglalás

Cikkünkben a Történeti Magánéleti Korpusz (TMK) webes lekérdezőfelületén elérhető interaktív tematikus-szemantikus lexikai térképet mutattuk be a kereső egyéb újdonságai mellett. A pusztán a TMK-ból készített, a korpusz kis mérete miatt jellegében inkább tematikusnak, mint igazán nyelvinek mondható szóbeágyazási modell mellett a TMK kibővítésével nyert korpuszból készített már inkább nyelvi-szemantikus modellekből a t-SNE algoritmussal nyert kétdimenziós lexikai térképek elemeire kattintva közvetlenül is indítható az adott nyelvi elemre vonatkozó korpuszlekérdezés. A térképek ugyanakkor a szövegek gépi feldolgozásakor, illetve kézi ellenőrzésekor bent maradt hibákra is felhívják a figyelmet, könnyítve ezzel a hibajavítást.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással a K 116217 számú projekt, illetve a K 15 pályázati program keretében valósult meg.

Hivatkozások

- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
- Dömötör, A., Gugán, K., Novák, A., Varga, M.: Kiútkeresés a morfológiai labirintusból : korpuszépítés ó- és középmagyar kori magánéleti szövegekből. *Nyelvtudományi Közlemények* 113, 87–114 (2017)
- Franz, M., Lopes, C.T., Huck, G., Dong, Y., Sümer, S.O., Bader, G.D.: Cytoscape.js: a graph theory library for visualisation and analysis. In: *Bioinformatics* (2015)
- van der Maaten, L., Hinton, G.E.: Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 9, 2579–2605 (2008)
- Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. *CoRR abs/1301.3781* (2013), <http://arxiv.org/abs/1301.3781>
- Novák, A.: Milyen a jó Humor? In: *I. Magyar Számítógépes Nyelvészeti Konferencia*. pp. 138–144. SZTE, Szeged (2003)
- Novák, A., Laki, L.J., Novák, B.: CBOW/A: módosított CBOW algoritmus annotált szövegekből készített vektortérmodellek létrehozására. In: *XV. Magyar Számítógépes Nyelvészeti Konferencia*. pp. 37–48 (2019)
- Novák, A., Novák, B.: Magyar szóbeágyazási modellek kézi kiértékelése. In: *XIV. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2018*. pp. 67–77 (2018)
- Novák, A., Novák, B.: Bu-Bor-éK: grafikus címkenormalizáló eszköz. In: *XVI. Magyar Számítógépes Nyelvészeti Konferencia*. pp. 303–312 (2020)
- Novák, A., Novák, B., Wenszky, N.: Szóbeágyazási modellek vizualizációjára és böngészésére szolgáló webes felület. In: *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*. pp. 355–362 (2017)
- Novák, A., Wenszky, N.: Ó- és középmagyar szóalaktani elemző. In: *IX. Magyar Számítógépes Nyelvészeti Konferencia [Ninth Hungarian Conference on Computational Linguistics]*. pp. 170–181 (2013)
- Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013)*. pp. 539–545. Incoma Ltd. Shoumen, Bulgaria, Hissar, Bulgaria (2013)
- Petersen, U.: Emdros — a text database engine for analyzed or annotated text. In: *In: Proceedings of COLING 2004*. (2004) 1190–1193 (2004)

BESZÉDTECHNOLÓGIA

3D konvolúciós neuronhálón és neurális vokóderen alapuló némabeszéd-interfész

Tóth László¹, Amin Honarmandi Shandiz¹, Gosztolya Gábor², Zainkó Csaba³,
Markó Alexandra^{4,5}, Csapó Tamás Gábor^{3,4}

¹Szegedi Tudományegyetem, Informatikai Intézet

²MTA-SZTE Mesterséges Intelligencia Kutatócsoport

³Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és
Médiainformatikai Tanszék

⁴MTA-ELTE „Lendület” Lingvális Artikuláció Kutatócsoport

⁵Eötvös Loránd Tudományegyetem, Alkalmazott Nyelvészeti és Fonetikai Tanszék
{tothl,shandiz,ggabor}@inf.u-szeged.hu, {csapot, zainko}@tmit.bme.hu,
marko.alexandra@btk.elte.hu

Kivonat A némabeszéd-interfészek célja beszédjel előállításával, az artikulációs szervek mozgását rögzítő felvételtől, például a nyelvmozgást tartalmazó ultrahang-videóból. Jelenleg erre a konverzióra a mély neuronhálókat alkalmazó megoldások tűnnek a legígéretesebbnek. Képek felismerésére már régóta alkalmazzák a konvolúciós neuronhálókat, a legjobb eredményt azonban akkor kaphatjuk, ha a videó egyes képkockáit nem külön-külön, hanem sorozatként dolgozzuk fel. Egy lehetséges megoldás erre, ha a képeket feldolgozó konvolúciós háló kimeneteinek sorozatát egy visszacsatolt neuronhálózattal egyesítjük. Jelen cikkben viszont egy másik megoldással próbálkozunk, nevezetesen 3-dimenziós konvolúciós hálókat használunk, ahol a képek két dimenziója mellett az idő képezi a harmadik tengelyt. A 3D konvolúciós hálónak is egy speciális változatát alkalmazzuk, amely a térbeli és időbeli konvolúciós lépéseket felbontott formában végzi el – ezt a fajta hálózatot sikeresen használták már más videófelismerési feladatokban is. Kísérleteinkben a 3D neuronháló némileg pontosabb eredményeket adott, mint a kombinált konvolúciós+visszacsatolt modell, ami azt mutatja, hogy ez a megközelítés alternatívája lehet a rekurrens hálókra épülő, általában lassabban és nehezebben tanítható modelleknek.

Kulcsszavak: némabeszéd-interfész, CNN, 3D konvolúció, neurális vokóder, nyelv-ultrahang

1. Bevezetés

Az utóbbi években megnőtt az érdeklődés az ún. „articulatory-to-acoustic” átalakítás iránt, amelynek célja az elhangzott beszéd becslése, visszaállítása pusztán az artikulációs szervek mozgása alapján. Ennek a leképezésnek a megoldása tudná a technológiai hátteret nyújtani olyan alkalmazások számára, mint például a némabeszéd-interfész (Silent Speech Interface, SSI (Denby és mtsai, 2010;

Schultz és mtsai, 2017)). Az artikulációs szervek mozgásának valamilyen felvétele alapján ugyanis elvileg olyan esetben is meg lehetne becsülni a beszédjelet, amikor az alany valójában nem is ejt ki hangot, azaz „némán beszél”. Az ilyen némabeszéd-interfészek segítségével vissza tudnánk adni olyan betegek beszéd-készségét, akik artikulációs szerveiket ugyan képesek mozgatni, de a hangadás képességét elvesztették (például a gégejük, hangszalagjaikat érintő műtét vagy sérülés következtében). Emellett olyan alkalmazási területek is felmerülnek, amikor a hangos beszédkommunikáció valamilyen más okból nem lehetséges (pl. nagyon zajos környezetben vagy bizonyos katonai alkalmazásokban). Az artikulációs szervek mozgásának követésére többféle megoldás létezik, a legegyszerűbb (bár nem teljes értékű) ezek közül az ajkak mozgásának videóra rögzítése (Ephrat és Peleg, 2017; Akbari és mtsai, 2018). További lehetőségként kínálkozik az elektromágneses artikulográfia (electromagnetic articulography (EMA), Kim és mtsai (2017a,b)), az ultrahangos nyelvkövetés (ultrasound tongue imaging (UTI), Jaumard-Hakoun és mtsai (2016); Csapó és mtsai (2017); Grósz és mtsai (2018); Kimura és mtsai (2019)) az állandó mágnessel készült artikulográfia (permanent magnetic articulography (PMA), Gonzalez és mtsai (2017)). Az artikulációs izmok elektromiográfiás figyelése (surface electromyography (sEMG), Maier-Hein és mtsai (2005); Janke és mtsai (2012); Janke és Diener (2017)) is lehetséges, illetve több szerző a fenti módszerek párhuzamos, kombinált használatával próbálkozik (Denby és mtsai, 2010). Mi ebben a cikkben a nyelvmozgásról készült ultrahang-videókból fogunk kiindulni.

Az artikulációs mozgásról készült felvételek beszédjellé konvertálásának konvencionálisabb módja a kétlépéses, felismerésből majd szintézisből álló eljárás (Schultz és mtsai, 2017). Mint a neve is mutatja, ez a megközelítés a rögzített jel alapján első lépésben megkísérli felismerni az elhangzott beszédet, majd a felismert szövegből beszéd szintézis útján állítja elő a beszédjelet (Denby és mtsai, 2011; Hueber és mtsai, 2010; Wang és mtsai, 2014). Ennek a módszernek a fő hátránya, hogy egyrészt nagy késleltetés keletkezhet az input és az output között, másrészt pedig a beszéd felismerő hibái zavaró módon megjelennek a szintetizált beszédben. További probléma, hogy az esetleges prozódiai információt teljesen elveszítjük az eredeti jelből, pedig a fő prozódiai komponensek – szünetek, hang-erő, sőt még az alapfrekvencia is – egész jól rekonstruálhatóak az artikulációs felvételtől (Grósz és mtsai, 2018).

A fentiek miatt a jelenlegi SSI megoldások a közvetlen szintézis elvét preferálják, azaz az artikulációs jelet közvetlenül beszédjellé próbálják alakítani, bármiféle közbülső lépés nélkül. A közvetlen leképezés elfogadható minőségű megoldását a mély neuronhálós (Deep Neural Network, DNN) technológia elterjedése tette lehetővé. Az „articulatory-to-acoustic” leképezéssel próbálkozó legújabb cikkek mindegyike mély neuronhálós technológiát alkalmaz, bármilyen jeltörzítési módszerről legyen is szó (Csapó és mtsai, 2017; Grósz és mtsai, 2018; Janke és Diener, 2017; Jaumard-Hakoun és mtsai, 2016; Gonzalez és mtsai, 2017; Moliner és Csapó, 2019; Kimura és mtsai, 2019; Saha és mtsai, 2020).

Jelen cikkben mély neuronhálókat fogunk alkalmazni a nyelvmozgást rögzítő ultrahang-videók beszédjellé alakítására. Habár az ugyanezen problémával fog-

lalkozó legkorábbi cikkek a legegyszerűbb felépítésű, ún. teljes kapcsolású (fully connected) neuronhálókat alkalmazták (Jaumard-Hakoun és mtsai, 2016; Csapó és mtsai, 2017), mivel a videóink képekből állnak, ésszerűbbnek ígérkezik konvolúciós neuronhálókat (convolutional neural network, CNN) használni. A konvolúciós hálók rendkívül sikeresek a kép alakfelismerésben (Krizhevsky és mtsai, 2012), és számos újabb, SSI-vel foglalkozó tanulmány már konvolúciós hálókra épül (Janke és Diener, 2017; Moliner és Csapó, 2019; Kimura és mtsai, 2019).

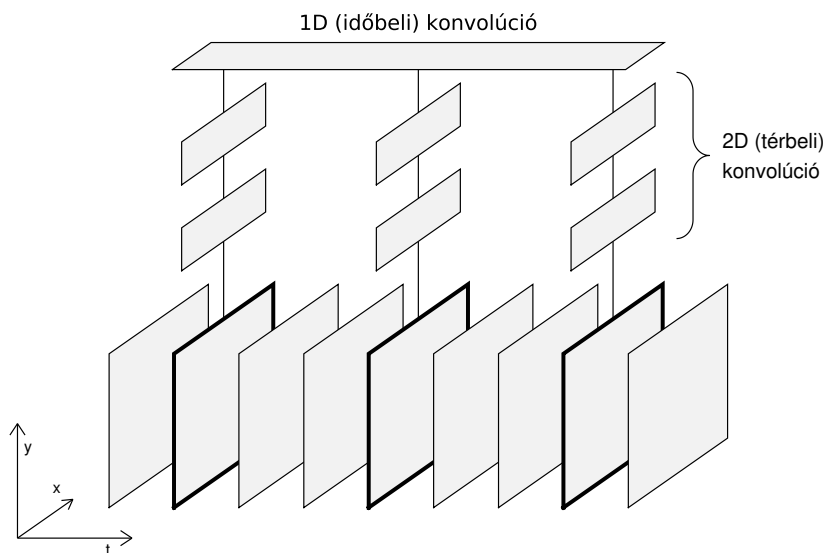
Esetünkben azonban az input egy videófelvétel, azaz nem egyetlen állókép, hanem képek sorozata. Egyetlen statikus képhez viszonyítva a képek sorozata nyilván extra információt tartalmaz a nyelvmozgás irányáról, dinamikájáról, tehát érdemes lehet egyidejűleg több szomszédos képből álló blokkokat feldolgozni. Sorozatok feldolgozására több neuronhálós struktúra is szóba jöhet. Ilyen esetben tipikusan visszacsatolt neuronhálókat szoktunk alkalmazni, például „long short-term memory”, röviden LSTM neuronhálót. Mivel az input képekből áll, ezért sztemendnek mondható megoldás az egyes képeket egy (kétdimenziós) konvolúciós hálóval feldolgozni, majd az egyedi képekből kinyert információt egy LSTM hálóval összevonni (Gonzalez és mtsai, 2017; Moliner és Csapó, 2019; Liu és mtsai, 2018; Kim és mtsai, 2017b). Erre a sémára röviden „CNN+LSTM” modellként fogunk hivatkozni.

Habár az LSTM hálók jól beváltak, folyamatos kritika éri őket a betanításuk lassúsága és nehézsége miatt. Mivel a visszacsatoló kapcsolatok miatt tanításuk nehezen párhuzamosítható, folyamatos a kutatás a hasonló hatásfokkal működő, de tisztán előreccsatolt hálózatok iránt, például a gépi fordítás (Lakew és mtsai, 2018) vagy a beszédfelismerés (Dong és mtsai, 2018) terén. Esetünkben ilyen alternatívaként merül fel a kétdimenziós konvolúciós háló (2D CNN) kiterjesztése három dimenzióra, ahol a két térbeli kiterjedés mellé az időtengety adja a harmadik dimenziót (Ji és mtsai, 2013; Kimura és mtsai, 2019; Wu és mtsai, 2018). A cikkben ezzel a megoldással fogunk kísérletezni, méghozzá egy speciális 3D hálóstruktúrát, ún. „(2+1)D CNN”-t használva (Tran és mtsai, 2018). Mint látni fogjuk, az eredmények azt mutatják, hogy ez a háló a visszacsatolt hálókéval ekvivalens eredményeket tud elérni ugyanannyi paraméter és rövidebb tanítási idő mellett. Bár további alapos összehasonlító vizsgálatokra lesz szükség, azt mindenképpen kimondhatjuk, hogy az ultrahang-videón alapuló néma-beszéd-interfészek esetén a 3D-konvolúcióra épülő előreccsatolt hálózatok a visszacsatolt LSTM-hálózatok életképes alternatívájának tűnnek.

A cikk a következőképp épül fel. A 2. fejezet az alkalmazandó konvolúciós háló technikai részleteit mutatja be. A 3. fejezetben ismertetjük a beszédjeleken és ultrahang-videókon alkalmazott adatkinyerési és -feldolgozási lépéseket, a 4. fejezetben pedig a kísérletek paramétereit. Az 5. fejezet az eredményeket mutatja be és vitatja meg, majd a cikket a levont konklúziókkal zárjuk a 6. fejezetben.

2. Videók feldolgozása konvolúciós neuronhálókkal

Az 'Alexnet' hálózat feltalálása óta az állóképek felismerésében a konvolúciós háló számít a legjobb technológiának (Krizhevsky és mtsai, 2012). Ezek a CNN



1. ábra: A $(2+1)$ D CNN háló szerkezetének illusztrációja. A videó képkockáit (alul) először 2D konvolúciót alkalmazó neurális rétegek dolgozzák fel, majd ezek kimenetét egy 1D konvolúciót végző réteg fogja össze. A modell átugorhat képkockákat; az ugrás lépésköze az időbeli konvolúció „stride” paraméterével szabályozható.

hálózatok a konvolúciót két dimenzióban, a kép két térbeli kiterjedése mentén végzik. Számos alkalmazásban azonban az input egy videó, nem pedig egyetlen állókép. Ilyen esetben a képkockákat sorozatként feldolgozva (külön-külön feldolgozás helyett) gyakran jelentősen javíthatók a felismerési eredmények. A legjobb példa erre az emberi járásmód (gait) felismerése, de általánosságban felhozható bármilyen mozgási esemény detektálása (action recognition, Ji és mtsai (2013); Zhao és mtsai (2018a,b)). Ezekben az esetekben a képkockák sorozata háromdimenziós tömbként fogható fel, ahol a két térbeli dimenzió mellé az idő adja a harmadik dimenziót (lásd 1. ábra).

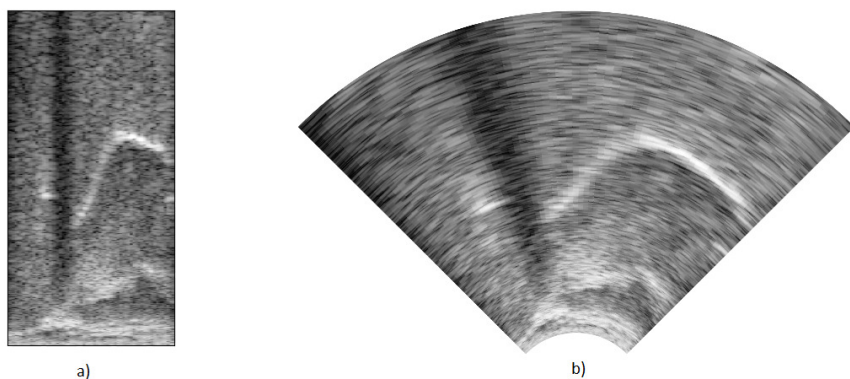
Sorozatok feldolgozásában hagyományosan a visszacsatolt neuronhálók, például az LSTM hálók számítanak a leghatékonyabb, legcélravezetőbb hálózattípusnak (Hochreiter és Schmidhuber, 1997). Ezen hálók betanítása azonban közismerten nehézkes és lassú, ezért több egyszerűsített változatuk is készült, mint például a GRU (gated recurrent unit (Cho és mtsai, 2014)) vagy a „kvázirekurrens” háló (Bradbury és mtsai, 2017). Emellett folyamatos kutatás folyik olyan struktúrák iránt, amelyek hasonló pontosságot tudnak elérni, de visszacsatolt kapcsolatok nélkül. Beszédfelismerésben például az időkésleltetési háló (time-delay neural network, TDNN) bizonyult kifejezetten sikeresnek (Peddinti és mtsai, 2015; Tóth, 2014), de megemlíthetjük az előrecsatolt szekvenciális memóriájú hálót is (feedforward sequential memory network, FSMN) (Zhang és mtsai, 2018). Videók felismerésére pedig a CNN struktúra többféle, sorozat-

tok feldolgozására alkalmas kiterjesztését javasolták (Ji és mtsai, 2013; Zhao és mtsai, 2018a,b). Sajnos a szokványos, 2D konvolúciót sokféleképp lehet kiterjeszteni három dimenzióra, így az optimális modell megtalálása nem triviális. Tran és munkatársai kísérleti úton vetettek össze többféle 3D hálózariánst, és a legjobb eredményeket a térbeli és időbeli konvolúciós lépések szétválasztásával kapták (Tran és mtsai, 2018). Az általuk „(2+1)D” CNN-nek nevezett modell először 2D konvolúciót végez a két térbeli tengely mentén, ezt követi az 1D konvolúció az időtengely mentén (lásd 1. ábra). Az 1D konvolúció lépésközparaméterének (stride) állításával átugorhatunk képkockákat, lényegében alulmintavételezve őket. Ennek értelme, hogy ily módon – persze rosszabb felbontással, de – szélesebb időtartományból nyerhetünk ki információt anélkül, hogy a hálózat méretét növelnünk kellene. Látni fogjuk, hogy ennek a tényezőnek kulcsszerepe lesz a jó eredmény elérésében, valamint érdekességképp megjegyezzük, hogy lényegében ugyanez az alapötlet a beszédfelismerésben is rendkívül hatékonynak bizonyult (Tóth, 2014). Természetesen ilyen (2+1)D blokkokból többet is egymásra pakolhatunk, ily módon mély hálózatot építve (Tran és mtsai, 2018). Videók felismerésére Luo és munkatársai is hasonló struktúrát találtak optimálisnak (Luo és Yuille, 2019), és az időbeli információ hasonló, hierarchikus összevonása történik a beszédfelismerésben népszerű TDNN hálóban is (Peddinti és mtsai, 2015).

3. Neurális vokóderek

Bár léteznek már olyan neurális struktúrák, amelyek kimenetként beszédjelet állítanak elő (például épp az itt bemutatandó neurális vokóderek), ezek többnyire nagyméretű hálózatok, amelyek betanításához nagyságrendileg nagyobb mennyiségű hanganyag szükséges, mint ami nekünk rendelkezésünkre állt. A kevés adat miatt ezért célravezetőbbnek ígérkezett közvetlenül a beszédjel helyett valamilyen nagyon tömör reprezentációt használni a tanulás kimeneteként. Természetesen ennek a reprezentációnak olyannak kell lennie, hogy abból a beszédjel aztán jó minőségben visszaállítható legyen. E célra kézenfekvően kínálta magát a beszédkódolásban használatos vokóderek által kinyert spektrális reprezentáció: ez egyrészt tömör (hiszen a cél a beszéd-tömörítés), másrészt visszaállítható belőle a beszéd (hiszen a betömörített beszédet ki is kell tudni tömöríteni).

A nyelvfeldolgozás számos más területe mellett a neuronhálók a beszédkódolásban és a beszéd-szintézisben is megjelentek. A neuronhálós beszéd-szintézisben az alábbi kétlépéses eljárás terjedt el: első lépésben a szöveget valamilyen spektrális reprezentációvá, például spektrogrammá vagy mel-spektrogrammá alakítják, a második lépésben pedig a becsült spektrogramból megkapják a beszédjelet (Prenger és mtsai, 2019). Esetünkben az első komponens szerepét a 3D konvolúciós háló fogja átvenni, hiszen az inputunk nem szöveg, hanem egy videó. A második komponens viszont – tkp. a vokódert – minden változtatás nélkül használni tudjuk. Az utóbbi időben számos, céljainkra használható neurális vokóder született. Mi ezek közül a WaveGlow-t választottuk (Prenger és mtsai, 2019), mivel a korábbiaknál egyszerűbben használhatónak és jobb minőségűnek tűnt,



2. ábra: Az ultrahang-felvételek megjelenítése a) a nyers adattömbnek megfelelő négyzetes elrendezésben b) interpolációval előállítható, anatómiaiilag korrekt elrendezésben.

illetve előre betanított modell is rendelkezésre állt hozzá. Habár ez a betanított modell angol nyelvű, korábbi lehallgatásos kísérletek azt mutatták, hogy magyar nyelvű beszéd szintetizálására is remekül használható – olyannyira, hogy a modell magyar nyelvű mintákon való újratanítása sem eredményezett lényegesen jobb minőségű magyar szintetizált beszédet. (Csapó és mtsai, 2020).

3.1. Adatrögzítés és -feldolgozás

Az ultrahang-felvételek egy magyar anyanyelvű női adatközlő közreműködésével készültek. A mondatok felolvasása során nyelvének mozgását az álla alatt rögzített ultrahang-fejjel vettük fel, az Articulate Instruments Ltd. „Micro” fantázianevű eszközrendszerét használva. Ez a berendezés másodpercenként 82 képet készít, mellyel párhuzamosan a beszédjelet is rögzítettük egy Audio-Technica ATR 3350 típusú kondenzátormikrofonnal, melyet a beszélő előtt 20 cm-re helyeztünk el. Az ultrahang-videó és a beszédjel szinkronizálására a berendezéshez tartozó szoftvert használtuk. Összesen 438 mondatot (körülbelül fél órányi hanganyagot) vettünk fel, melyet véletlenszerűen osztottunk tanító, validációs és teszthalmazokra 310-41-87 arányban. Ugyanezt az adatbázist már korábbi tanulmányok is használták (Csapó és mtsai, 2017; Grósz és mtsai, 2018).

A berendezés a 64 pásztázó nyaláb mindegyike mentén 946 mintát rögzít, amelyből a megfelelő szoftverekkel végzett interpoláció útján az anatómiai viszonyoknak megfelelő ábrát kaphatunk (lásd 2. ábra jobb oldala). Azonban ez a fajta ábra a szokatlan alakja miatt nehezebben feldolgozható, miközben nem tartalmaz extra információt az eredeti 946x64 méretű adattömbhöz képest (2. ábra, bal oldal). Emiatt közvetlenül a nyers adatokkal dolgoztunk, sőt, az adattömböt újramintavételezéssel 128x64 méretűre kicsinyítettük. A képek intenzitásértékeit a $[-1, 1]$ intervallumra normalizáltuk.

A beszédjelet 22050 Hz mintavételezéssel rögzítettük, majd egy, a beszéd-szintézisben is használt függvény-implementációval mel-frekvenciaskálás spektrogrammá konvertáltuk. A feldolgozás lépésközt 270 mintára állítottuk, ez felelt meg legjobban az ultrahang 82 kép/sec felbontásának. Mivel a WaveGlow betanításakor 256-os lépésközt használtak (ez 86,1 kép/sec-et jelent), az eltérésből eredő enyhe elcsúszást a szintézis előtt a mel-spektrogram interpolációjával korrigáltuk. A WaveGlow inputjaként szolgáló mel-spektrogram a 0-8000 Hz-es frekvenciatartományt 80 sávra osztja, így neuronhálónk tanítása során ez a 80 komponensből álló spektrális vektor szolgált tanítási célértékként.

4. Kísérleti konfigurációk

Neuronhálónkat a Keras keretrendszer segítségével, Tensorflow alapon implementáltuk (Chollet és mtsai, 2015). Öt különböző modellt készítettünk: egy egyetlen képkockát feldolgozó teljes kapcsolású hálót (FCN), egy 2D konvolúciós hálót, amely továbbra is egyetlen képkockán dolgozik (2D CNN), valamint egy 3D konvolúciós hálót, amely már képkockák sorozatát kapja bemenetként (3D CNN); összehasonlítási alapként készítettünk egy hálót, amely a képkockákat feldolgozó 2D-CNN rétegek eredményét egy LSTM segítségével összegzi (CNN+LSTM), valamint ebből a hálóból készítettünk egy kétirányú változatot is (CNN+BiLSTM). Hogy paraméterszám tekintetében összemérhetőek legyenek, mindegyik hálót úgy lőttük be, hogy kb. 4,3 millió tanulható paraméterük legyen. A tanításra Adam optimalizálót használtunk 100-as batchmérettel. Tanítási hibafüggvényként az átlagos négyzetes hibát (mean squared error, MSE) alkalmaztuk.

Teljes kapcsolású háló (FCN): A lehető legegyszerűbb hálóstruktúra, ha teljes kapcsolású (Keras nyelven „Dense”) rétegeket használunk. Esetünkben a hálót öt, rétegenként 430 neuront tartalmazó rejtett rétegből építettük fel, míg a kimeneti réteg 80 darab, lineáris aktivációjú neuronból állt, a mel-spektrális célvektornak megfelelően. A háló inputja egyetlen képkocka, azaz $128 \times 64 = 8192$ pixel. A rejtett rétegek a swish aktivációs függvényt használták (Ramachandran és mtsai, 2017), és minden rejtett réteg után egy dropout réteg következett 0,2-es kiejtési valószínűséggel.

Konvolúciós neuronháló (2D CNN): Az előző hálóhoz hasonlóan ez a háló is egyetlen képkockát dolgoz fel, azonban a legfelső, Dense réteg alatti négy réteg mindegyike térbeli konvolúciót végez az adatokon. A rétegek részletes konfigurációja az 1. táblázatból olvasható le. A legjobbnak tűnő meta-paramétereket kísérletezgetéssel kerestük meg, a rejtett rétegek ebben az esetben is a swish aktivációs függvényt alkalmazták.

3D Konvolúciós neuronháló (3D CNN): Ebben a hálózatban a 2D konvolúció helyett 3D konvolúciót alkalmaztunk, mivel ez teszi lehetővé egyetlen képkocka helyett képek rövid sorozatának feldolgozását. A konkrét hálóstruktúra (lásd 1. táblázat) öt képkockát dolgoz fel, amelyek s távolságra találhatók egymástól, ahol s az időtengely mentén végzett konvolúció lépésköze („stride” paramétere). A 2. fejezetben bemutatott (2+1)D konvolúció koncepciójának meg-

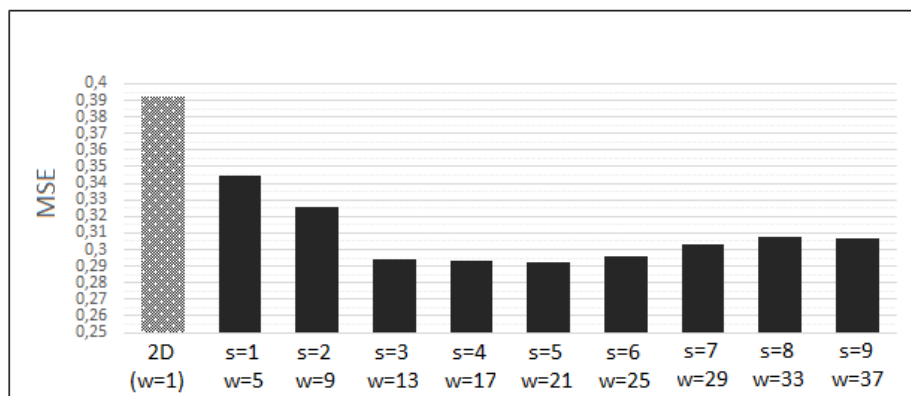
1. táblázat. A 2D és a 3D CNN hálók rétegei a Keras-implementációban, valamint a legfontosabb paramétereik. A különbségeket vastag betűvel kiemeltük.

2D CNN	3D CNN
Conv2D(30, (13,13), strides=(2,2))	Conv 3D (30, (5 ,13,13), strides=(s ,2,2))
Dropout(0.2)	Dropout(0.2)
Conv2D(60, (13,13), strides=(2,2))	Conv 3D (60, (1 ,13,13), strides=(1 ,2,2))
Dropout(0.2)	Dropout(0.2)
MaxPooling2D(pool_size=(2,2))	MaxPooling 3D (pool_size=(1 ,2,2))
Conv2D(90, (13,13), strides=(2,1))	Conv 3D (90, (1 ,13,13), strides=(1 ,2,1))
Dropout(0.2)	Dropout(0.2)
Conv2D(150, (13,13), strides=(2,2))	Conv 3D (85 , (1 ,13,13), strides=(1 ,2,2))
Dropout(0.2)	Dropout(0.2)
MaxPooling2D(pool_size=(2,2))	MaxPooling 3D (pool_size=(1 ,2,2))
Flatten()	Flatten()
Dense(1000)	Dense(1000)
Dropout(0.2)	Dropout(0.2)
Dense(13, activation='linear')	Dense(13, activation='linear')

felelően az öt képkockát először külön-külön dolgozza fel a háló, majd ezután végzi el az időtengely menti konvolúciós lépést. Az 1. táblázatban vastag betűvel emeltük ki, hogy ehhez milyen módosításokat kellett eszközölni az eredeti, 2D CNN hálóhoz képest. Megjegyezzük, hogy a legfelső konvolúciós réteg méretének csökkentése azért volt szükséges, hogy a két háló paraméterszáma megközelítőleg ugyanannyi maradjon.

LSTM neuronháló (CNN+LSTM): A teljes kapcsolású, illetve a 2D CNN hálók egyetlen képkockát dolgoznak fel, így sejthető, hogy nem lesznek méltó versenytársai a képek sorozatán működő 3D CNN hálónak. Sorozatok feldolgozására az LSTM hálózat ajánlott, illetve mivel képek sorozatáról van szó, érdemes a visszacsatolt hálót a képeket feldolgozó 2D CNN hálóval kombinálni. Az általunk használt CNN+LSTM modell alsó négy rétege megegyezett a 2D CNN háló alsó négy rétegével; legfelső rejtett réteggként viszont a Dense réteget LSTM rétegre cseréltük. A paraméterszám megőrzésre miatt 500 LSTM neuront tettünk ebbe rétegbe, a hálózat inputját pedig 21 egymást követő képkocka képezte.

Kétirányú LSTM neuronháló (CNN+BiLSTM): Ha nem ragaszkodunk a teljesen valós idejű feldolgozáshoz, akkor az LSTM réteg nem csak időben előre (balról jobbra), hanem időben visszafelé (jobbról balra) haladva is működhet. Szokásos megoldás továbbá egy előre és egy visszafelé haladó réteget is képezni, és ezek kimenetét kombinálni. A CNN+BiLSTM nevű modellünk egy ilyen, úgynevezett kétirányú (bidirectional) hálót takar, amely az előző, CNN+LSTM modelltől csakis az LSTM réteg kétirányúsításában tér el. A paraméterszám megőrzése miatt ebben a modellben az LSTM réteg méretét 320-ra kellett redukálnunk.



3. ábra: 3D CNN háló MSE hibaértéke a validációs halmazon az s paraméter függvényében. Összehasonlításképp a 2D CNN háló hibáját is feltüntettük (bal szélső oszlop).

5. Kísérleti eredmények

A neuronhálóink teljesítményének kiértékelésére többféle lehetőség kínálkozik. A legegyszerűbb megoldás a tanítás során használt célfüggvény (jelen esetünkben az MSE) értékeit összehasonlítani a validációs vagy a tesztadatokon. Az MSE értékek mellett a korreláción alapuló R^2 értékeket is fel fogjuk tüntetni, mivel regressziós feladatok esetén az is egy egyszerűen kiszámolható és népszerű mérőszám. Azonban mivel a kimenetünk egy beszédjel, ezek a matematikai alapon megfogalmazott egyszerű mérőszámok nem feltétlenül tükrözik a hang érzékelt, szubjektív minőségét. Erre vonatkozóan csak lehallgatásos tesztekkel – mint például a MUSHRA teszt (ITU, 2001) – kaphatnánk becslést. Az ilyen, emberi alanyokkal történő szubjektív kiértékelés azonban elég fáradságos, ezért dolgozták ki a különféle objektív metrikákat, amelyek ugyan matematikai úton, de az emberi hallás fő tulajdonságait figyelembe véve próbálják megbecsülni a hang minőségét. Kísérleteink első részében csak a két egyszerű, objektív mérőszámot – a MSE ill. R^2 értékeket – közöljük, a végső összehasonlításnál azonban néhány további, hangminőséget becslő objektív mérőszámot is mutatunk majd.

Mint az elméleti ismertetésben láttuk, a 3D konvolúciós hálónak van egy nagyon fontos meta-paramétere, az s paraméter. A legelső kísérletben ennek hatását vizsgáltuk a hibafüggvény értékére. Ez a paraméter határozza meg, hogy a háló az input mekkora időszakaszáról kap információt: a két szélső képkocka közötti távolság a $w = 4 \cdot s + 1$ képlettel határozható meg. Például $s = 5$ érték esetén a háló által lefedett időablak mérete $w = 21$ képkocka. A videó 82 kép/sec mintavételezési rátáját figyelembe véve, ez körülbelül negyed másodpercnak, nagyságrendileg egy szótag hosszának felel meg.

2. táblázat. A különféle hálóarchitektúrákkal kapott MSE és R^2 értékek a validációs és a teszhalmazon.

Hálózat típusa	Val		Teszt	
	MSE	Mean R^2	MSE	Mean R^2
FCN	0,410	0,600	0,419	0,585
2D CNN	0,392	0,617	0,401	0,603
3D CNN ($s=5$)	0,292	0,714	0,293	0,710
CNN + LSTM	0,303	0,701	0,296	0,709
CNN + BiLSTM	0,301	0,706	0,296	0,707

A 3. ábra mutatja a 3D CNN hálóval kapott MSE értékeket az s paraméter különböző értéke esetén. Összehasonlításképp a 2D CNN háló (amely csak egyetlen képkockát dolgoz fel) hibáját is feltüntettük. Látható, hogy az aktuális képkocka mellett annak környezetét is figyelembe véve jelentős hibacsökkentést érhetünk el. Már 2-2 közvetlen szomszédot használva ($s=1$) is jobb eredményt kapunk, de lényegesen nagyobb a javulás 3-6 közti s értékekkel. A tágabb kontextus figyelembe vétele tehát fontos, még képkockák átugrása árán is. A fenti eredmény alapján s értékét 5-re rögzítettük.

A következő kísérletben az ötféle hálóstruktúrát vetettük össze, a validációs és teszhalmazokon kapott MSE és R^2 értékeket a 2. táblázat összegzi (R^2 esetén a nagyobb érték jelent jobb modellt). Látható, hogy az egyetlen képkockát feldolgozó FC és 2D CNN hálók közül a konvolúciós háló ugyan egyértelműen jobb, de sokkal jobb eredményt érhetünk el az egyetlen kép helyett képsorokat feldolgozó hálóvariánsokkal (3D CNN ill. CNN+LSTM hálók).

A 3D CNN és az LSTM-alapú hálók összevetéséhez az LSTM hálók inputját 21 képkockára állítottuk, hiszen a 3D CNN esetén ez bizonyult optimálisnak. Mint a táblázatból látható, a CNN+LSTM modell egyértelműen megverte ugyan az egyetlen képkockás modelleket, de a 3D CNN háló pontosságát nem tudta Meghaladni. Nem változtatott ezen az LSTM réteg kétirányúsítása (BiLSTM) sem: míg ez más feladatokon általában egy pici javulást szokott hozni, itt most gyakorlatilag az egyirányú hálóval ekvivalens eredményt kaptunk. Felvetődött, hogy esetleg a CNN+LSTM modellek számára más lehet az optimális ablakméret, ezért próbáltunk változtatni a 21-es inputméreten, de más értékek esetén sem nem kaptunk lényegesen jobb eredményt. A kapott hibaértékek alapján úgy tűnik, hogy a képkockák alulmintevételezése ugyanolyan hatékonyan segíti az információ fúzióját, mint az összes képkocka feldolgozása az LSTM szofisztikáltabb, visszacsatolást és belső memóriát is alkalmazó technikájával. Az LSTM viszont, épp a rekurrens jellege miatt, nem tud átugrani képkockákat, pedig lehet, hogy ebben az esetben pont erre lenne szükség. Szintén az összes képkocka megőrzéséből kifolyólag az CNN+LSTM háló tanítása jóval hosszabb – kb. 70%-kal több – időt vett igénybe, mint a 3D CNN háló betanítása. A modellek azonos paraméterszáma ellenére érdekes módon az LSTM háló memóriáigénye is nagyobb volt, ennek feltehetően szintén az összes input-képkocka megőrzése az oka.

3. táblázat. Öt modellünk összevetése beszédminőséget mérő objektív mérőszámokkal.

	STOI	PESQ	MCD
FCN	0,661	1,562	4,602
2D CNN	0,658	1,551	4,569
3D CNN (s=5)	0,743	1,831	4,161
CNN + LSTM	0,742	1,792	4,139
CNN + BiLSTM	0,736	1,789	4,152

Végezetül megjegyezzük, hogy korábban Moliner és Csapó is próbálkozott a 2D CNN és LSTM hálók kombinálásával hasonló feladaton (Moliner és Csapó, 2019). Eredményük azonban direkt módon nem összevethető a miénkkel, ugyanis másik vokódert, és ennek megfelelően a tanítás során más célértékeket használtak. Emellett az általuk használt háló jóval nagyobb is volt, több mint négyszeres paraméterszámmal a mi hálóinkhoz képest. Az egyirányú és a kétirányú LSTM-változatok teljesítménye között ők sem tapasztaltak szignifikáns különbséget. Vélünk párhuzamosan Saha és munkatársai is próbálkoztak nyelvtultrahang-videók feldolgozásával, és tőlünk függetlenül a miénkhez hasonló 3D konvolúciós hálóstruktúrát hoztak ki optimálisnak, valamint ők lényegesen jobb eredményeket kaptak a 3D CNN hálóval, mint a CNN+LSTM kombinációval (Saha és mtsai, 2020).

A hang minőségének kiértékelésére sokféle objektív mérőszámot javasoltak. Ezek valamilyen szinten igyekeznek figyelembe venni az emberi hallás működésének fő tulajdonságait, így valamivel pontosabb becslést adnak a hangminőségre, mint a tanítás során optimalizált MSE hibafüggvény. Az alábbi, 3. táblázatban három ilyen mérőszámot értékeltünk ki az ötféle modellel szintetizált tesztalmazon, ezek sorban a STOI (Short-Term Objective Intelligibility, Taal és mtsai (2011)), a PESQ (Perceptual Evaluation of Speech Quality, ITU-R (2001)), valamint az MCD (Mel-Cepstral Distance, Kubichek (1993)). Előbbi kettő esetén a magasabb érték jelez jobb minőséget, utóbbi esetén a kisebb. Az így kapott a számok is at mutatják, hogy egyértelmű minőségi ugrás van az egyetlen képkeockát feldolgozó (FC és 2D CNN), valamint a képsorozatokat konvertáló (3D CNN, LSTM és BiLSTM) hálók között. Két mérőszám a 3D hálót, egy pedig az LSTM hálót hozta ki győztesnek, de az eltérés e két háló teljesítménye között mindhárom metrika szerint minimális, így egyértelmű nyertest nem mernénk hirdetni.

6. Összefoglaló

Cikkünkben egy háromdimenziós konvolúciót végző neuronháló hatékonyságát vizsgáltuk a beszédjel artikulációs felvételekből való visszaállításának feladatkorében. Tran és munkatársainak tanulmánya által motiválva, hálózatunk a 3D konvolúciót két lépésre bontja, így előbbi a térbeli, majd az időbeli konvolúciós lépést végzi el (Tran és mtsai, 2018). Modellünket egy másik, képsorozatokat

modellezésére javasolt hálótípussal, a CNN+LSTM háló két változatával vetettük össze. Kísérleteinkben a 3D háló ekvivalens, vagy kicsivel jobb teljesítményt nyújtott, miközben betanítása kevesebb időt igényelt. Mivel mindkét modellnek sok meta-paramétere van, a 2D CNN fölényének kijelentéséhez további alapos mérések kellenének, de annyit bizonyosan állíthatunk, hogy a 3D CNN háló mindenképpen versenyképes alternatívát jelent a CNN+LSTM kombinált hálókkal szemben, ha a célunk nyelvultrahang-videókon alapuló néma beszéd-interfész építése. A jövőben további összehasonlításokat tervezünk egy újabb hálótípussal, az úgynevezett ConvLSTM hálókkal. Mint nevük is mutatja, ezek a háló egyetlen rétegen belül egyesítik a konvolúciós és az LSTM modellezés előnyeit, tehát elvileg még alkalmasabbak lehetnek videók feldolgozására (Zhao és mtsai, 2019).

Köszönetnyilvánítás

A kutatást a Nemzeti Kutatási Fejlesztési és Innovációs Hivatal FK 124584 kódjelű pályázata, valamint az Innovációs és Technológiai Minisztérium TUD-FO/47138-1/2019-ITM programja támogatta. Gosztolya Gábor kutatásait az MTA Bolyai János kutatói ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosító: ÚNKP-20-5-SZTE-649) támogatta. A kutatáshoz használt grafikus kártya az NVIDIA Corporation ajándéka.

Hivatkozások

- Akbari, H., Arora, H., Cao, L., Mesgarani, N.: LIP2AUDSPEC : Speech reconstruction from silent lip movements video. In: Proc. ICASSP. pp. 2516–2520 (2018)
- Bradbury, J., Merity, S., Xiong, C., Socher, R.: Quasi-recurrent neural networks. In: Proc. ICLR (2017)
- Cho, K., van Merriënboer, B., Gülcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: Proc. EMNLP. pp. 1724–1734 (2014)
- Chollet, F., és mtsai: Keras. <https://github.com/fchollet/keras> (2015)
- Csapó, T.G., Grósz, T., Gosztolya, G., Tóth, L., Markó, A.: DNN-based ultrasound-to-speech conversion for a silent speech interface. In: Proc. Interspeech. pp. 3672–3676 (2017)
- Csapó, T.G., Zainkó, C., Tóth, L., Gosztolya, G., Markó, A.: Ultrasound-Based Articulatory-to-Acoustic Mapping with WaveGlow Speech Synthesis. In: Proc. Interspeech 2020. pp. 2727–2731 (2020), <http://dx.doi.org/10.21437/Interspeech.2020-1031>
- Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Communication 52(4), 270–287 (2010)
- Denby, B., Cai, J., Hueber, T., Roussel, P., Dreyfus, G., Crevier-Buchman, L., Pillot-Loiseau, C., Chollet, G., Manitsaris, S., Stone, M.: Towards a practical silent speech interface based on vocal tract imaging. In: Proc. ISSP. pp. 89–94 (2011)

- Dong, L., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: Proc. ICASSP. pp. 5884–5888 (2018)
- Ephrat, A., Peleg, S.: Vid2speech: Speech reconstruction from silent video. In: Proc. ICASSP. pp. 5095–5099 (2017)
- Gonzalez, J.A., Cheah, L.A., Gomez, A.M., Green, P.D., Gilbert, J.M., Ell, S.R., Moore, R.K., Holdsworth, E.: Direct speech reconstruction from articulatory sensor data by machine learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25(12), 2362–2374 (2017)
- Grósz, T., Gosztolya, G., Tóth, L., Csapó, T.G., Markó, A.: F0 estimation for DNN-based ultrasound silent speech interfaces. In: Proc. ICASSP. pp. 291–295 (2018)
- Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* 9(8), 1735–1780 (1997)
- Hueber, T., Benaroya, E.L., Chollet, G., Dreyfus, G., Stone, M.: Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication* 52(4), 288–300 (2010)
- ITU: ITU-R recommendation BS.1534: Method for the subjective assessment of intermediate audio quality (2001)
- ITU-R: ITU-R recommendation P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (2001)
- Janke, M., Diener, L.: EMG-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25(12), 2375–2385 (2017)
- Janke, M., Wand, M., Nakamura, K., Schultz, T.: Further investigations on EMG-to-speech conversion. In: Proc. ICASSP. pp. 365–368 (2012)
- Jaumard-Hakoun, A., Xu, K., Leboullenger, C., Roussel-Ragot, P., Denby, B.: An articulatory-based singing voice synthesis using tongue and lips imaging. In: Proc. Interspeech. pp. 1467–1471 (2016)
- Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(1), 221–231 (2013)
- Kim, M., Cao, B., Mau, T., Wang, J.: Multiview representation learning via deep CCA for silent speech recognition. In: Proc. Interspeech. pp. 2769–2773 (2017a)
- Kim, M., Cao, B., Mau, T., Wang, J.: Speaker-Independent Silent Speech Recognition From Flesh-Point Articulatory Movements Using an LSTM Neural Network. *IEEE/ACM Trans. ASLP* 25(12), 2323–2336 (2017b)
- Kimura, N., Kono, M., Rekimoto, J.: Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In: Proc. of CHI Conf. on Human Factors in Computing Systems (2019)
- Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems* 25. pp. 1097–1105 (2012)

- Kubichek, R.: Mel-cepstral distance measure for objective speech quality assessment. In: Proc. ICASSP. pp. 125–128 (1993)
- Lakew, S., Cettolo, M., Federico, M.: A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In: Proc. COLING. pp. 641–652 (2018)
- Liu, Z.C., Ling, Z.H., Dai, L.R.: Articulatory-to-acoustic conversion using BLSTM-RNNs with augmented input representation. *Speech Communication* 99(2017), 161–172 (2018)
- Luo, C., Yuille, A.: Grouped spatial-temporal aggregation for efficient action recognition. In: Proc. International Conference on Computer Vision. pp. 5512–5521 (2019)
- Maier-Hein, L., Metze, F., Schultz, T., Waibel, A.: Session independent non-audible speech recognition using surface electromyography. In: Proc. ASRU. pp. 331–336 (2005)
- Moliner, E., Csapó, T.: Ultrasound-based silent speech interface using convolutional and recurrent neural networks. *Acta Acustica united with Acustica* 105 (2019)
- Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proc. Interspeech. pp. 3214–3218 (2015)
- Prenger, R., Valle, R., Catanzaro, B.: Waveglow: A flowbased generative network for speech synthesis. In: Proc. ICASSP. pp. 3617–3621 (2019)
- Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a Self-Gated Activation Function. *ArXiv e-prints* 1710.05941 (2017)
- Saha, P., Liu, Y., Gick, B., Fels, S.: Ultra2speech – a deep learning framework for formant frequency estimation and tracking from ultrasound tongue images. In: Proc. MICCAI. pp. 473–482 (2020)
- Schultz, T., Wand, M., Hueber, T., Krusienski, D.J., Herff, C., Brumberg, J.S.: Biosignal-based spoken communication: A survey. *IEEE/ACM Trans. ASLP* 25(12), 2257–2271 (2017)
- Taal, C., Hendriks, R., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Trans. ASLP* 19(7), 2125–2136 (2011)
- Tóth, L.: Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition. In: Proc. ICASSP. pp. 190–194 (2014)
- Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proc. CVPR (2018)
- Wang, J., Samal, A., Green, J.: Preliminary test of a real-time, interactive silent speech interface based on electromagnetic articulograph. In: Proc. SLPAT. pp. 38–45 (2014)
- Wu, C., Chen, S., Sheng, G., Roussel, P., Denby, B.: Predicting tongue motion in unlabeled ultrasound video using 3D convolutional neural networks. In: Proc. ICASSP. pp. 5764–5768 (2018)
- Zhang, S., Lei, M., Yan, Z., Dai, L.: Deep-FSMN for large vocabulary continuous speech recognition. In: Proc. ICASSP (2018)

- Zhao, C., Zhang, J., Wu, C., Wang, H., Xu, K.: Predicting tongue motion in unlabeled ultrasound video using convolutional LSTM neural networks. In: Proc. ICASSP. pp. 5926–5930 (2019)
- Zhao, S., Liu, Y., Han, Y., Hong, R., HU, Q., Tian, Q.: Pooling the convolutional layers in deep convnets for video action recognition. IEEE Trans. Circuits and Systems for Video Technology 28(8), 1839–1849 (2018a)
- Zhao, Y., Xiong, Y., Lin, D.: Trajectory convolution for action recognition. In: Advances in Neural Information Processing Systems 31. pp. 2204–2215 (2018b)

End-to-end és hibrid mélyneuronháló alapú gépi leiratozás magyar nyelvű telefonos ügyfélszolgálati beszélgetésekre

Mihajlik Péter^{1,2}, Balog András², Tarján Balázs^{1,3}, Fegyő Tibor^{1,3}

¹ Budapest Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformaticai Tanszék, 1111, Budapest, Műegyetem rkp. 3.
mihajlik@tmit.bme.hu

² THINKTech Nonprofit Kft., 2600, Vác, Váczy Pál u. 15.
abalog@thinktech.hu

³ SpeechTex Kft, 1181, Madách Imre u. 47.
{tarjan, fegyő}@speechtex.com

Kivonat A tisztán mélyneuronhálóra épülő gépi beszédfelismerés alig pár éve került a tudományos köztudatba, de máris az egyik leginkább kutatott szakterületté vált. Magyar nyelvre történő alkalmazása, illetve annak publikációja azonban ez idáig váratott magára. Cikkünkben beszámolunk az első tapasztalatainkról a téren, magyar nyelvű telefonos ügyfélszolgálati beszéd leiratozása témában. A valós idejű működés érdekében nagy számú, egy dimenziós, idő- és csatornatengely szerint szeparált konvolúciós réteget alkalmazunk reziduális kapcsolatokkal és normalizációkkal. Karakter akusztikus modellekkel, szótár és nyelvmodell nélkül is összevethető, bár gyengébb szó- és betűpontossági eredményeket kaptunk a hibrid (rejtett Markov-modell + mélyneuronháló) rendszeréhez képest. Nyelvmodell hozzáadásával és angol nyelven előtanított súlyokkal történő inicializálás alkalmazásával szignifikáns javulást mértünk, meghaladva a hibrid eredményeket. Kutatásunk megerősíti, hogy az end-to-end beszédfelismerési modellezés magyar nyelven is életképes, azonban a teljes potenciál eléréséhez valószínűleg nagyságrendi növekedésre lenne szükség az akusztikus tanítóadatok (hang + leirat) terén.

1 Bevezetés

Történelmi távlatból nézve a beszédfelismerés hatékonysága, használhatósága szempontjából mindig az adat (hullámforma + leirat) jelentette a kulcsot. A dinamikus idővetemítés sikere annak volt köszönhető, hogy intuitív frekvenciaelemzés helyett valós referenciabeszéd-felvételekhez hasonlította a felismerendő bemondást (Sakoe és Chiba, 1978). A beszélőfüggetlenséget a nagy mennyiségű beszédadatot felhasználni képes, már valódi gépi tanulás alapú rejtett Markov-modell (Jelinek és mtsai, 1975) tette lehetővé. A folyamatos beszéd szöveggé alakításában pedig annak felismerése volt a kulcslépés, hogy az egyes szószorozat-hipotézisekhez az a-priori valószínűségeket –

n-gram nyelvi modelleken keresztül – a témához illeszkedő nagy mennyiségű szöveg alapján becsülhetjük (Kunh és De Mori, 1990).

Ebbe a trendbe illeszkedik a mélytanulás: erősen leegyszerűsítve úgy is mondhatjuk, hogy a mély neuronhálók fő erénye a (sokkal) több adat (sokkal) hatékonyabb hasznosítása, melyhez persze némi célzott hardware támogatás is (GPU) szükségeltetik.

Miután mind az akusztikus, mind a nyelvi modellek terén a mélyneuronhálók alkalmazása vált egyeduralkodóvá a state-of-the-art rendszerekben, valamint a graféma alapú akusztikus modellek már korábban is jól teljesítettek (Mihajlik és mtsai, 2007), észszerű irányként merült fel a bemenő akusztikus megfigyelések közvetlenül neuronhálóval történő leképezése szó/karakter szekvenciává. Az end-to-end – azaz elejétől végéig mély neuronháló alapú – megközelítés nem hozott azonnali áttörést (Graves és Jaitly, 2014), inkább letisztultságával tűnt ki, azonban rövid idő alatt rendkívül kutatott és sikeres területté vált. Angol nyelv esetén a mai napig folyamatos versenyben vannak a hibrid, mély tanulást és HMM struktúrát is a használó, ill. a tisztán mélytanulási technikák – a cikk írása idején a Switchboard korpuszon éppen (Tüske és mtsai, 2020) érte el a legjobb eredményt egy viszonylag egyszerű end-to-end struktúrával, míg a LibriS-peechen (Pan és mtsai, 2020) vezet hibrid megközelítéssel.

Így indokoltnak láttuk egy releváns, telefonközpontos beszélgetéseket tartalmazó magyar nyelvű adatbázison megvizsgálni a két beszédfelismerési irányzat egy-egy képviselőjének eredményeit.

2 A beszédfelismerési feladat

A magyar nyelvű, ügyfélszolgálati témakörű telefonos beszéd felismerése kiemelt jelentőséggel bír napjainkban. Így a rendelkezésünkre álló összes, nem publikus, anonimizált telefonközponti leiratozott beszélgetést felhasználva alakítottuk ki a tanító, validáló és kiértékelő halmazokat az 1. táblázat szerint.

1. Táblázat: Telefonos ügyfélszolgálati beszédatadabázis részhalmazai

	Tanító	Validáló	Kiértékelő
Audio (óra)	290	7	12
Text (szószám)	3.4M	46K	66K

A beszélgetések valós, spontán, ügyfélszolgálati beszélgetéseket tartalmaznak, olykor jelentős háttérzajjal terheltek. A tanító, validáló és kiértékelő halmazokra bontásnál alapfeltétel volt, hogy egy beszélgetés csak egy halmazba kerülhet. Míg a validáló és kiértékelő halmazokba csak teljes, vágatlan beszélgetéseket tettünk, a tanító halmazból a beszélgetések egyes részei (pl. a túl zajos szakaszok, vagy a lejegyző által nem értett részek) kikerülhettek.

A beszédfelismerési folyamatra megkötés, hogy csak valós időben alkalmazható technológiák jöhetnek szóba (így pl. két körös beszédfelismerés, vagy BLSTM struktúra sem). További szempont az alacsony késleltetés és a nagy offline áteresztőképesség (egy feldolgozási időegység alatt minél több bejövő beszédidőegység feldolgozása).

2 Vizsgált megközelítések

2.1 TDNN-F – HMM hibrid akusztikus modell

Klasszikus rejtett Markov-modell alapú megközelítés, ahol környezetfüggő (bal difón) fonéma akusztikus modelleket (AM) használunk és az egyes HMM állapotokhoz tartozó valószínűség-sűrűség függvényeket modellezzük mély neuronhálóval. A GMM-es előtanítást követően, annak időszegmentálását felhasználva tanítjuk be a faktorált Time Delay Neural Network (Povey és mtsai, 2018) architektúrát (a továbbiakban: TDNN-F). A tanítás lépései és az alkalmazott módszerek megegyeznek a Kaldi¹ LibriSpeech s5 receptúrájával – azzal a megkötéssel, hogy az általános használhatóság kedvéért a beszélőcímkéket igénylő i-vektorokat nem alkalmazzuk. Bemeneti vektorokként nagy felbontású MFCC-t használtunk, az audio adatok dúsítására (augmentálására) a felvételek sebességét és hangerejét perturbáltuk. A TDNN-F paraméterszáma hozzávetőlegesen 18.5M volt.

2.2 Idő- és csatornatengely mentén szeparált 1D konvolúciós háló-alapú end-to-end akusztikus modell

Connectionist Temporal Classification (CTC) költségfüggvénnyel (Graves és mtsai, 2006) tanított, teljes kiépítésben 78 rétegű 1D konvolúciós neuronháló (Kriman és mtsai, 2020) karakterszintű kimenettel. A paramétertér csökkentése érdekében a konvolúció szétválik csatornánkénti FIR (Finite Impulse Response) szűrésre, majd a szűrt csatornák lineáris kombinációjára ill. a nemlinearitás alkalmazására. Normalizálásnak a batchnorm-ot használtuk, aktivációs függvénynek a ReLU-t. A gradiens visszaterjesztés elősegítésére az 5-ös blokkokra bontott konvolúciós rétegeket átívelő reziduális kapcsolatot alkalmaztunk. A reprodukálhatóság kedvéért a standard receptúrát (Kriman és mtsai, 2020) követtük itt is. A bemenő akusztikus adatok Mel-skálázott rövid idejű amplitúdó spektrumok voltak. SpecAugment-et (Park és mtsai, 2019) minden esetben, hangerő és beszédsebesség perturbációt opcionálisan alkalmaztunk. A mély-neuronháló össz. paraméterszáma 18.9M.

2.3 Nyelvmodellek

Alapértelmezésben hagyományos, szó alapú back-off 4-gram nyelvmodellt (LM: Language Model) alkalmaztunk a szokásos, módosított Kneser-Ney (Chen és Goodman, 1999) simitással. A magyar nyelvhez jobban illeszkedő morf nyelvi modellt is kipróbáltunk, további részletek a neurális tudástranszfer alapú nyelvmodellezésről a (Tarján és mtsai, 2020)-ban találhatóak – mi a jelen kutatásban a kisebb (1 GB) memóriaigényű augmentált morf modellt használtuk.

¹ <https://github.com/kaldi-asr/kaldi>

3 Kísérleti eredmények

3.1 Kísérleti elrendezés

End-to-end esetben lehetőség van tisztán akusztikai alapon történő leiratozásra. Ennek legegyszerűbb módszere a „greedy” algoritmus: CTC módszerrel tanított neurális modellek kimeneteiből keretszinkron módon a legvalószínűbb karaktert kiválasztjuk, a közvetlenül egymás után ismétlődőket egyetlen karakterrel helyettesítjük, majd a „blank” karaktert kiszedve összeolvassuk az eredményt. Természetesen a szóköz a tanításnál kötelező a szavak között, így a felismerési fázisban a szóhatárok természetes módon visszaállítódnak (helyes felismerés esetén).

HMM-es beszédfelismerésnél a súlyozott kiejtési alternatívákat, a fonemikus környezetfüggőséget és a nyelvmodellt WFST keretrendszerben (Mohri és mtsai, 2002) integráljuk és optimalizáljuk, majd a szokásos beam-search eljárással választjuk ki – szintén keretszinkron módon – a (Viterbi közelítéssel mért) legvalószínűbb szószorozat-hipotézist.

A korrekt összehasonlíthatóság érdekében a nyelvmodell end-to-end akusztikus modellel történő kombinációját a HMM-es elrendezéssel azonos módon, ugyanazon beam-search dekódolóval végezzük. Ekkor a karakter kiejtési szótár formális, csupán a „blank” karakterek beékelődésire kell felkészíteni, ill. nincs fonológiai értelemben vett környezetfüggés.

A dekódolást mindig telítésközel munkapontban végeztük, így a dekódolási sebesség a valós időnél még mindig kb. 70-szer gyorsabb. Ebbe az akusztikus hasonlóságot számoló neuronhálók „inference” számítását nem értjük bele, ami GTX 1080 TI GPU alkalmazásával a valós időtől kb. 60-szor gyorsabb.

Az optimális nyelvi és akusztikus modell súlyozást a validáló halmazon végeztük.

3.2 Kalibrációs tesztek

Az első lépés a kijelölt megközelítések ellenőrzésére, hogy ismert angol nyelvű beszéd-felismerési feladaton lemérjük a pontosságukat. A publikus LibriSpeech adatbázist Panayotov és mtsai, 2015) választottuk, ennek teljes tanítóanyagán (960 óra) tanítottunk mindkét esetben, illetve a standard „test clean” halmazon értékeltük ki az eredményeket.

A TDNN-F modelleket az egyik standard („3-gram ARPA LM, pruned with threshold $1e-7$ ”) nyelvi modellel értékeltük ki. A szófelismerési hiba (Word Error Rate) a 2. táblázatban látható, marginálisan – jobb, mint a Kaldi saját közlése (5.3%).

2. Táblázat: Kalibrációs eredmények a LibriSpeech (960 óra) adatbázison

AM	LM	WER (test clean)
TDNN-F	word	5.24%
end-to-end	–	5.20%
end-to-end	word	3.78%

Az end-to-end rendszert a rendelkezésre álló szűkösebb GPU memóriakapacitások miatt (4 x GTX 1080 TI / 11GB) kisebb batch mérettel (64) és kevesebb epoch számmal (200) tanítottuk, mint (Kriman és mtsai, 2020) ajánlják, azonban így is jobb eredményt ért el, mint a hibrid rendszer lexikon és nyelvmódel nélkül is, nyelvmódel hozzáadásával pedig előnye szignifikánsan megnőtt. Ezek ellenére, a (Kriman és mtsai, 2020) által közölt pontosságot (3.9% nyelvmódel nélkül) nem értük el, aminek oka a rövidebb tanítás, a beszédsebesség-perturbáció hiánya, valamint a kevésbé kimerítő hiperparaméter-optimalizálás lehetett. Mindazonáltal, a „kalibrációt” sikeresnek tekintettük, mind a hibrid, mind a tisztán neuronháló alapú megközelítés működik, hozzá az elvárt eredményeket.

3.3 Telefonos ügyfélszolgálati beszédfelismerési teszteredmények

A telefonos ügyfélszolgálati beszédadatok akusztikus modelltanításra előkészítése jelentette az első feladatot. Itt különféle előszegmentálások és szűrések után gyakorlatilag kereken 200 óra tanítóanyag maradt közvetlenül felhasználható a tanításra. Mindkét alább vizsgált megközelítésnél ugyanezt a tanítóanyagot használtuk és mindenhol közöljük a szóhibaarány (WER) mellett a betűhibaarányt (LER) is.

A TDNN-F modelleket az előzőek szerint tanítottuk, mindössze azzal a különbséggel, hogy a bemenetünk most 8kHz-es mintavételezésű, így az előfeldolgozást ennek megfelelően a Kaldi Switchboard receptúrája szerint (minimálisan) módosítottuk. A kiértékelő tesztalmazon mért beszédfelismerési eredmények a 3. táblázatban találhatók.

3. Táblázat: Beszédfelismerési eredmények
magyar nyelvű telefonos ügyfélszolgálati adatokon

AM	LM	WER	LER
TDNN-F	word	21.40%	9.93%
TDNN-F	morf	18.96%	9.19%
end-to-end (baseline)	–	30.63%	12.52%
end-to-end (augment)	–	29.07%	12.71%
end-to-end (pretrain)	–	28.54%	11.81%
end-to-end (pretrain + la)	–	27.65%	12.36%
end-to-end (pretrain + la + augment)	–	26.07%	11.84%
end-to-end (pretrain + la + augment)	word	18.79%	9.56%
end-to-end (pretrain + la + augment)	morf	17.83%	9.15%

Az end-to-end modellek esetén első körben szintén 8kHz-es mintavételezésre állítottuk a megfelelő alacsony szintű jelfeldolgozási paramétert és csak a learning-rate értékét optimalizáltuk, valamint a 200 epoch-os tanítási hosszát (egy hetes futásidő az ismertetett hardveren) megtartottuk (baseline). Ezután – a mindig bekapcsolt „on-the-fly” spectral augment mellett – beszédsebesség és hangerő perturbációt alkalmaztunk (augment). Majd 16kHz-es mintavételezést/adatkonvertálást beállítva, az angol nyelvű LibriSpeech-en az előző pontban betanított neurális hálózat súlyaival inicializáltuk a

tanítást (csak az encoder hálózatot). Így drasztikusan csökkentett, 15-ös epoch szám mellett is az előzőeknél jobb eredményt kaptunk (pretrain), amit az alacsony szintű beszédjellemzők betanulásának és az így megvalósult tudásátadásnak (transfer learning) tulajdonítunk. Végül hosszabb, 45-ös epoch esetén („la”, mint long adaptation) további javulást tapasztaltunk (pretrain + la), valamint a korábbi augmentáció és hosszabb, angol nyelven előtanított súlyokkal inicializált tanítás esetén még további javulást (pretrain + la + augment).

Ahogy az 3. táblázat mutatja, a (greedy) end-to-end eredmények – főleg szóhibarány tekintetében – némiképp elmaradnak a TDNN-F rendszeréhez képest. Ugyanakkor, összehasonlítva egy korábbi MSZNY konferencián bemutatott rendszerünk eredményével, ahol előreccsatolt DNN ACM-et alkalmaztunk (Tarján és mtsai, 2019), az end-to-end modellek úgy mutatnak javulást, hogy se kiejtési szótárt, se nyelvmoddell nem használnak.

Nyelvmoddell hozzáadásával az end-to-end szóhibarányok ugrásszerűen javulnak, a betűhibarányok – a felismerési elvből fakadóan – érthetően kevésbé. Viszont mind WER, mind LER tekintetében sikerült áttörést elérni: az azonos nyelvmoddellel mért hibrid és end-to-end eredmények közül minden esetben az utóbbiak bizonyultak jobbak. Ezzel együtt is, az end-to-end beszédfelismerési eredmények a hibrid TDNN-F rendszeréhez viszonyítva kisebb javulást hoztak a vártnál. Ennek elsődleges okaként a „kalibrációs” adathalmazhoz képesti jóval kisebb tanítóadatmértetet tudjuk megjelölni.

Megjegyezzük, hogy az itt publikált hibrid mély-neuronhálós eredmények ugyan számszerűen a legalacsonyabbak, de a TDNN-F hibrid rendszer telefonos magyar ügyfélszolgálati nyelvre alkalmazása nem új, csupán a kiértékelésre használt referencialeírat tisztítása okozza a látszólagos javulást (Tarján és mtsai, 2020)-hoz képest.

4 Következtetések

Megvizsgáltuk, hogy a nagy áteresztőképességgel működő mélytanulásra épülő beszédfelismerési módszerek milyen pontosságot érhetnek el magyar nyelvű telefonos ügyfélszolgálati beszéd gépi leiratozásánál. Azt kaptuk, hogy a rendelkezésre álló akusztikus tanítóadatbázis-méret mellett az angol nyelven széles körben alkalmazott end-to-end megközelítés nyelvmoddell alkalmazása nélkül is összevethető eredményeket ad a jelenleg legjobb hibrid megközelítéshez képest, a neurális tudástranszferrel készült morf nyelvmoddell alkalmazásával pedig felül is múlja azt. Meggyőződésünk, hogy elsősorban az akusztikus tanítóadat mennyiségét szükséges növelni a még jobb, az angol nyelvű eredményekkel összemérhető pontosságértékekért. A nemzetközi eredményekkel összevetésben és jelen vizsgálatok alapján is úgy látjuk, jelenleg nem a HMM struktúra létén/nem létén múlik a pontosság, hanem sokkal inkább az alkalmazott mélytanulási módszereken. A magyar nyelv digitális fenntarthatósága érdekében tehát elsősorban a megfelelő tanítóadatok (hangfelvétel + leirat) nagyságrendi növelése lenne a cél, másodsorban pedig a kiszolgáló számítástechnikai infrastruktúra (korszerű GPU gridek) fejlesztése, például, hogy a hiperparaméterek érdemi optimalizálására is legyen reális lehetőség.

Hivatkozások

- Chen, S. F. and Goodman, J.: "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393 (1999)
- Graves, A.; Fernández, S.; Gomez, F., Schmidhuber, J.: "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". *ICML 2006*, pp. 369–376, (2006)
- Graves, A. and Jaitly, N.: "Towards End-To-End Speech Recognition with Recurrent Neural Networks." *ICML (2014)*.
- Jelinek, F.; Bahl, L.; Mercer, R.: "Design of a linguistic statistical decoder for the recognition of continuous speech". *IEEE Transactions on Information Theory*. 21 (3), pp. 250. (1975).
- Kriman S. et al., "Quartznet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, , pp. 6124-6128, (2020)
- Kuhn, R. and De Mori, R.: "A cache-based natural language model for speech recognition." *IEEE Transactions on pattern analysis and machine intelligence* 12.6: 570-583(1990)
- Mihajlik, P., Fegyő, T., Tüske Z., and Ircing P.: "A Morpho-graphemic Approach for the Recognition of Spontaneous Speech in Agglutinative Languages – like Hungarian," *Interspeech'07*, Antwerp, Belgium, (2007)
- Mohri, M, Pereira, F. and Riley, M.: "Weighted Finite-State Transducers in Speech Recognition", *Computer Speech and Language*, 16(1), pp. 69–88, (2002)
- Pan J., Shapiro J., Wohlwend J., Han K. J., Lei T., and Ma T., "ASAPP-ASR: Multistream CNN and Self-Attentive SRU for SOTA Speech Recognition," in *Proc. INTERSPEECH*, pp. 16–20. (2020)
- Panayotov V., Chen G., Povey D., and Khudanpur S., "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, pp. 5206–5210 (2015)
- Park, D. S. et al., "SpecAugment: A simple data augmentation method for automatic speech recognition", in *Proc. Interspeech*, (2019)
- Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S.: Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks. *Proc. Interspeech*, 3743-3747. (2018)
- Sakoe, H.; Chiba, S.: "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 26 (1), pp. 43–49. (1978)
- Tarján, B, Fegyő, T és Mihajlik, P.: Ügyfélszolgálati beszélgetések nyelvmodellezése rekurens neurális hálózatokkal, in *Proc MSZNY (2019)*
- Tarján, B, Szaszák G, Fegyő T, Mihajlik P: Improving Real-time Recognition of Morphologically Rich Speech with Transformer Language Model, in *Proc 11th IEEE International Conference on Cognitive Infocommunications (2020)*
- Tüske, Z; Saon, G; Audhkhasi, K; Kingsbury, B.: Single headed attention based sequence-to-sequence model for state-of-the-art results on Switchboard, in *Proc Interspeech (2020)*

Enyhe kognitív zavar detektálása beszédhangból x-vektor reprezentáció használatával

José Vicente Egas-López¹, Balogh Réka², Imre Nóra², Tóth László¹,
Vincze Veronika³, Pákáski Magdolna², Kálmán János²,
Hoffmann Ildikó⁴, Gosztolya Gábor^{1,3}

¹ Szegedi Tudományegyetem, Informatikai Intézet

² Szegedi Tudományegyetem, Pszichiátriai Klinika

³ MTA-SZTE Mesterséges Intelligencia Kutatócsoport, Szeged

⁴ Nyelvtudományi Intézet, Budapest

{ egasj, ggabor } @ inf.u-szeged.hu

Kivonat Az enyhe kognitív zavar (EKZ) heterogén klinikai szindróma, melyet gyakran tartanak a demencia preklinikai (azaz a demencia diagnózis felállításához nem elegendő, de mérhető kognitív hanyatlással járó) szakaszának is. Az EKZ jellemzői közé tartozik a kognitív funkciók enyhe hanyatlása, beleértve a memóriát, a végrehajtó és a nyelvi funkciókat. Kutatások alapján a nyelvi funkciók megváltozása már azelőtt észlelhető, hogy az EKZ-ra jellemző egyéb kognitív tünetek megjelenjenek. Az alanyok beszédének elemzése így praktikus, olcsó és nem-invazív eszköze lehetne a betegség korai szűrésének. Jelen munkánkban egy viszonylag friss, mély neurális hálón alapuló eljárást, az x-vektorokat használjuk jellemzőkinyerésre, majd ezen jellemzőket felhasználva osztályozó eljárást (SVM-et) tanítunk az EKZ-s és a kontroll beszélők elkülönítésére. Kísérleti eredményeink alapján az x-vektorokkal pontosabb diszkrimináció érhető el, mint a hagyományos i-vektorok használatával.

Kulcsszavak: demencia, enyhe kognitív zavar, x-vektorok

1. Bevezetés

A demencia krónikus, progresszív klinikai szindróma, amely főként idős személyeket érint világszerte. Jellemzői közé tartozik a memória, a nyelvi készségek és a problémamegoldó képesség romlása. A fenti készségeket érintő hanyatlás olyan mértékű, hogy az megnehezítheti vagy ellehetetlenítheti a páciens mindennapi tevékenységeinek elvégzését (Alzheimer’s Association, 2020). A betegség jelenleg kb. 46,8 millió embert érint világszerte, ez a szám pedig a becslések szerint 2050-re megduplázódhat (Prince és mtsai, 2015). Tekintve, hogy a jelenleg elérhető terápiás beavatkozások a betegség korai szakaszában vagy a betegséget megelőző, preklinikai stádiumban mutatják a legnagyobb hatékonyságot (Szatlóczi és mtsai, 2015), a betegség ezen fázisokban történő, korai felismerése kiemelt fontosságú.

A demencia preklinikai szakaszát a szakirodalom enyhe kognitív zavarnak (EKZ) nevezi. Ez az állapot egyfajta határterületnek tekinthető az öregedéshez

társuló, normális mértékűnek tekinthető szellemi hanyatlás és a már kimutatható demencia között (Petersen és mtsai, 2014). Számos kutatási eredmény utal arra, hogy az EKZ a páciensek beszédkésztségére is hatással van – ezekre támaszkodva a beszédelemzés költséghatékony, non-invazív eszközt kínálhat a betegség korai felismerésére. Az utóbbi években számos kutatás jelent meg, olyan eszközök és eljárások bemutatásával, amelyek célja egészséges kogníciójú kontroll (K) személyek és EKZ-val vagy Alzheimer-kórral élő páciensek automatizált módszerrel történő megkülönböztetése volt az alanyok beszédének vizsgálata alapján (lásd de Ipiña és mtsai, 2018; König és mtsai, 2018; Themistocleous és mtsai, 2018; Sluis és mtsai, 2020; Themistocleous és mtsai, 2020).

A szakirodalomban ismertetett eljárások egy részében feladat-specifikus jellemzőket vizsgáltak: olyan paramétereket kerestek tehát, amelyek eltérnek a kontrollszemélyek és az EKZ-s vagy enyhe AK-s alanyok beszédében. Ilyen paraméterek voltak például a szünetek száma és időtartama (Vincze és mtsai, 2020), vagy a beszédtempó és az artikulációs ráta (Meilán és mtsai, 2020). (A jellemzőkinyerést azután természetesen egy standard gépi tanulási lépés követi, például Support Vector Machine-t (SVM) használva.) Egy másik elterjedt megközelítés az, hogy *általános célú* eljárásokat alkalmazva nyernek ki jellemzőket az egyes alanyok hangfelvételeiből. Ezt követően ezeket a jellemzővektorokat felhasználva, statisztikai alapú osztályozó eljárással lehet elkülöníteni a két (vagy esetenként több) beszélőcsoportot. Ilyen általános célú jellemzővektorok lehetnek például az *i*-vektorok: habár ezeket eredetileg beszélőfelismerés céljára fejlesztették ki, később sikerrel alkalmazták a Parkinson-kór (García és mtsai, 2017; García és mtsai, 2018) és az Alzheimer-kór (Weiner és Schultz, 2018; Egas-López és mtsai, 2019) felismerésére is.

A beszélőfelismerés területén a korábban a legkorszerűbb technikának tartott *i*-vektorok helyét az utóbbi években egy mély neurális hálóra (Deep Neural Network, DNN) épülő eljárás, az *x*-vektorok vették át (Snyder és mtsai, 2018). A mély tanulás térhódítását tekintve ez nem is meglepő. Ésszerűnek tűnhet, hogy az *i*-vektorok után az *x*-vektorokat is alkalmazni kezdik az orvostudományi beszédfeldolgozás területén, vagy a technikai értelemben valamennyire rokon témakörnek számító paralingvisztikai feladatok esetén. Eddig ugyanakkor elég kevés ilyen tanulmány jelent meg: orvostudományi területen csak Botelho és munkatársai, valamint Jeancolas és munkatársai tanulmányairól van tudomásunk. Mindkét fent említett kutatócsoport a Parkinson-kór felismerésére alkalmazott *x*-vektorokat (Botelho és mtsai, 2020; Jeancolas és mtsai, 2020) (és mindkét tanulmány csak arXiv preprintként érhető el jelenleg).

Jelen cikkünkben azt vizsgáljuk, hogy milyen hatékonysággal alkalmazhatóak az *x*-vektor beágyazások az EKZ fölismerésére. Snyder és munkatársai egy előre tanított, letölthető modellt (DNN-t) is a közösség rendelkezésére bocsátottak; emellett a cikkben saját modellel is kísérletezünk, 60 órányi magyar nyelvű spontán beszédre tanítva. Az *x*-vektorok a háló több rétegéből is kinyerhetők, melyek hatékonyságát szintén megvizsgáljuk, az elért pontosságértékeket pedig összevetjük az *i*-vektorok használatával elért eredményekkel.

Réteg	Réteg közvetlen környezete	Teljes környezet mérete	Be- és kimenetek száma
Keret #1	$[t-2, t+2]$	5	120, 512
Keret #2	$\{t-2, t, t+2\}$	9	1536, 512
Keret #3	$\{t-3, t, t+3\}$	15	1536, 512
Keret #4	$\{t\}$	15	512, 512
Keret #5	$\{t\}$	15	512, 1500
Összegző	$[0, T]$	T	$1500T$, 3000
Szegmens #6	$\{0\}$	T	3000, 512
Szegmens #7	$\{0\}$	T	512, 512
Szoftmax	$\{0\}$	T	512, N

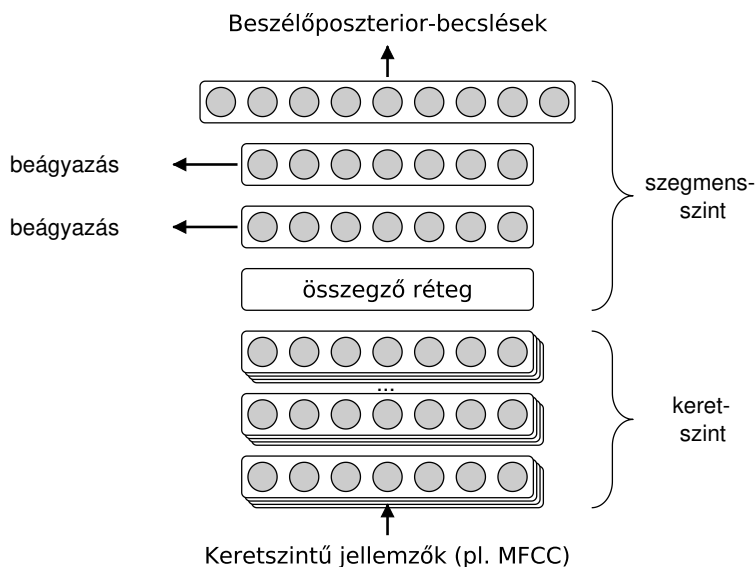
1. táblázat. Az x-vektor rendszer DNN architektúrája, mely öt keretszintű rétegből, egy statisztikai összegző (pooling) rétegből, két szegmensszintű rétegből és egy kimeneti, szoftmax rétegből áll. N -nel jelöltük a háló betanítására használt adatbázis beszélőinek számát (azaz a szoftmax réteg neuronjainak számát). Jelen architektúra megegyezik a Snyder és munkatársai által leírttal (Snyder és mtsai, 2017).

2. X-vektor kódolás

Az x-vektor technika egy neurálisháló-alapú jellemzőkinyerő eljárás, mely a változó hosszú hangfelvételeket fix dimenziószámú jellemzőtérbe képi. Technikailag egy mély neurális hálóról van szó, melynek bemenetei keretszintű vektorok (például MFCC-k), mélyebben elhelyezkedő rejtett rétegei keret-, magasabb rejtett rétegei pedig szegmensszinten működnek. Az egyes bemondásokhoz tartozó beágyazásokat (azaz az x-vektorokat) a szegmensszintű rétegek aktivációi jelentik.

A legelterjedtebb x-vektor architektúrát Snyder és munkatársai vezették be (Snyder és mtsai, 2018). Ebben a keretszintű rejtett rétegek időeltolások (*time-delay*) módon működnek: például a második keretszintű rejtett réteg t . kerethez tartozó aktivációinak meghatározásához az első keretszintű réteg három kerethez tartozó aktivációját ($t-2$, t és $t+2$) használjuk bemenetként. (Ld. 1. táblázat.) Az ötödik keretszintű réteg után egy összegző réteg (*stats pooling layer*) következik: ennek bemenete az utolsó keretszintű réteg *az aktuális felvétel összes keretén számítva*. (A 1. táblázatban a felvétel kereteinek számát T -vel jelöltük.) Az összegző réteg ezen aktivációk átlagát és szórását számolja ki; ezen két, 1500-1500 elemű vektor egymás után fűzve adja az első szegmensszintű réteg bemenetét. Ezen ponttól kezdve a háló szegmensszintűként működik tovább. A kimeneti, szoftmax réteg a tanító halmazban található beszélők számának megfelelő számú neuront tartalmaz (Snyder és mtsai, 2017, 2018).

A háló tanítása, a főnti struktúrát kihasználva, nem keret-, hanem szegmensszinten történik; címkeként az adott felvétel beszélőjének azonosítóját használjuk (mondjuk keresztentrópia veszteségfüggvénnyel). Tanítás után a beágyazások kinyerésére praktikusán bármelyik szegmensszintű réteg alkalmas; a tapasztalatok alapján a hatodik (a kimeneti rétegtől távolabb eső) réteg aktivációinak használata jobb eredményekhez vezet, mint a hetedik rétegé (Snyder és mtsai, 2018).



1. ábra: Az x-vektort kinyerő mély neurális háló általános struktúrája (Snyder és mtsai, 2018, nyomán).

Megjegyeznénk még, hogy a kimeneti réteg kizárólag tanításkor kap szerepet, így a későbbiekben ez el is dobható.

3. Kísérleti körülmények

3.1. Az EKZ-s és kontroll alanyok felvételei

A EKZ felismerésére vonatkozó kísérleteinket saját hangadatbázison végeztük, melyet a Szegedi Tudományegyetem Pszichiátriai Klinikáján rögzítettünk. A rögzítés digitális diktafonnal történt, külső mikrofon használatával; a felvételeket utólag monó, 16 kHz-es mintavételezésű formátumra konvertáltuk. Az alanyok spontán beszédét rögzítettük, az instrukció a következő volt: „Kérem, részletesen mesélje el az előző napját!”. (A felvételekről további részletekért ld. Vincze és mtsai, 2020). Az elkészült felvételekből hangminőség alapján válogattunk; jelen tanulmányunkban 25 EKZ-s és 25 kontroll alany felvételeit használtuk fel. A két csoport életkorbeli, nembeli és (elvégzett iskolai években mért) képzettségbeli eloszlása nem mutatott statisztikailag szignifikáns különbséget. Sajnos a felvételi körülmények miatt sok bemondás még a válogatás ellenére is visszhangos vagy jelentős háttérzajjal rendelkezett; a jel-zaj-arány (Signal-to-Noise Ratio, SNR) 14 és 35 dB közé esett.

3.2. Keretszintű jellemzők

Keretszintű jellemzőkészletként standard MFCC vektorokat használtunk. 20 MFCC együttthatót nyertünk ki a felvételekből 25 milliszekundum hosszú keretből, 10

milliszekundumos lépésközzel, a Kaldi eszköz segítségével (Povey és mtsai, 2011), melyhez hozzátettük még a lokális energiát mint jellemzőt. Bár terveink között szerepelt az egyéb típusú (keretszintű) jellemzők kipróbálása, az eszköz korlátai miatt nem volt lehetőségünk sem frekvencia-szűrősorok energiaértékeinek („filter banks”) használatára, sem az első- vagy másodrendű deriváltak felhasználására.

3.3. Az x-vektor DNN-ek tanítása

Saját x-vektor-kinyerő neurális háló modelljeinket a BEA Spontánbeszéd-adatbázis egy részhalmazán tanítottuk (Neuberger és mtsai, 2014). 165 beszélőt választottunk ki; a felvételekből automatikusan kivágtuk azokat a részeket, melyekben a felvételvezető hangja is hallható, így 10636 hangfelvételt kaptunk, összesen körülbelül 60 órányi terjedelemben. Az eredeti sztereó, 44,1 kHz-en mintavételezett bemondásokat monó, 16 kHz-es formátumra konvertáltuk.

A DNN modelleket a Kaldi rendszerrel (Povey és mtsai, 2011) tanítottuk be, és ezt használtuk a jellemzővektorok kinyerésére is. A tanítás során szokásos eljárás a tanító adat méretét mesterségesen megnövelni úgy, hogy az eredeti hangfelvételekhez zajt adnak és/vagy visszhangosítják azokat (Snyder és mtsai, 2018). Mivel az EKZ-s és kontroll alanyainktól gyűjtött hangfelvételeink eleve elég rossz minőségűek voltak, míg a BEA adatbázis szinte stúdióminőségű felvételeket tartalmaz, kíváncsiak voltunk, hogy ez a fajta augmentáció segíti-e az osztályozási lépést. Emiatt két DNN modellt tanítottunk: egyet augmentációval, egyet pedig ennek a lépésnek a kihagyásával. (Az augmentáció 52636 felvétellel (293 órányira) növelte a tanítóanyag méretét.)

A fenti két saját modell mellett kipróbáltunk egy előre betanított, szabadon elérhető modellt is (Snyder és mtsai, 2018). Ez a modell a Switchboard 2 Phases 1, 2 és 3, a Switchboard Cellular, valamint a NIST SRE adatbázisokon lett tanítva (összesen kb. hétezer beszélőn). További kisebb eltérés, hogy ez a modell 23 dimenziós MFCC-n (plusz az energián) működik, míg az általunk tanítottak, az i-vektoroknál szokásos méretet követve, 20 dimenzióson (korábbi tesztjeinkben azonban nem találtunk különbséget a 20 és a 23 dimenziós MFCC-t használó modellek között).

3.4. Jellemzőkinyerés

A beszélőtípusok azonosítására jellemzővektorként használt x-vektorokat a fent ismertetett három DNN modellből nyertük ki. Az 1. táblázatban leírt struktúrájú DNN-ből három ponton nyerhető ki felvételszintű vektor: az általában használt *Szegmens #6* réteg mellett a *Szegmens #7* és az *Összegző* réteg is alkalmas arra, hogy aktivációit (felvételszintű) jellemzőként használjuk. (Ezekben a rétegekben sorban 512, 512 és 3000 neuron található, így ezeknek megfelelő méretű jellemzővektort kapunk.) Mivel kíváncsiak voltunk, hogy enyhe kognitív zavar detektálására melyik réteg a legalkalmasabb, kísérleteinkben összesen kilenc variációt (három DNN modell és három réteg) próbáltunk ki. Emellett viszonyítási alapként i-vektorokat is használtunk jellemzőkként (128 komponens

alkalmazva); az i -vektorok általános háttérmodellje (Universal Background Model, UBM) az összehasonlíthatóság érdekében szintén a BEA adatbázis 3.3. fejezetben bemutatott részhalmazán lett tanítva. Az i -vektorok kiszámítására is a Kaldi rendszert használtuk.

3.5. Beszélőosztályozás

A jellemzőkinyerési lépés után a beszélőket Support Vector Machine (SVM, Schölkopf és mtsai, 2001) alkalmazásával, ötszörös keresztvalidációval osztályoztuk, a Python scikit-learn csomagját (Pedregosa és mtsai, 2011) használva. Minden SVM modell tanítása 20 EKZ-s és 20 kontroll alany hangfelvételén történt. Kiértékelési metrikáink a következők voltak: osztályozási pontosság (classification accuracy, *Pont.*), pontosság (precision, *Prec.*), fedés (recall), F_1 -érték (F -measure), valamint a ROC görbe alatti terület (AUC). (Pontosság (precision), fedés és F_1 esetén az EKZ beszélőkategóriát tekintettük pozitív osztálynak; mivel csak két beszélőkategóriánk (EKZ és kontroll) volt, a két osztályra kapott AUC-értékek megegyeztek.) A túltanulás elkerülése érdekében lineáris kernelt használtunk, így egyetlen hiperparaméterünk az SVM C (complexity) értéke volt; ezt az 10^{-5} , 10^{-4} , \dots , 10^1 értékeket végigpróbálva (grid search), a legnagyobb AUC értéket megcélözva választottuk ki. Előzetes tesztjeink eredményeit követve az x -vektorok esetén nem volt szükség a vektorok standardizálására vagy normalizálására, míg az i -vektorokat standardizáltuk (azaz minden jellemzőt nulla átlagra és egység szórásra transzformáltunk).

4. Eredmények

A 2. táblázat tartalmazza a különböző DNN modellekből és rétegekből kinyert x -vektor jellemzőket használva kapott pontosságértékeket. Látható, hogy a három használt DNN-réteg közül mindig az irodalomban általában ajánlott *Szegmens #6*-os réteg használatával kinyert jellemzőkkel kaptuk a legjobb eredményeket. Ennek oka feltehetőleg az, hogy az összegző réteg még nem foglalja össze a keretszintű információkat elég precízen, míg az utolsó rejtett réteg (*Szegmens #7*) már túlságosan feladat-specifikus információkat tárol (azaz túl specifikus a tanítóhalmazban szereplő beszélőkre).

A BEA adatbázison tanított két DNN modell közül az augmentálás használata valamivel jobb eredményekhez vezetett. Ez feltehetőleg annak köszönhető, hogy az augmentálási lépés amellett, hogy megnöveli a tanítóadat mennyiségét, zajtűrőbbé is teszi a modellt (mivel az extra tanítóadat az eredeti felvételek zajosított, illetve visszhangosított változataiból áll), amely hasznosnak bizonyulhat, amennyiben az osztályozandó felvételek nem éppen ideális körülmények között lettek rögzítve. Figyelembe véve, hogy a legtöbb beszédtechnológiai alkalmazás esetén nem várhatunk el stúdióminőséget, a modell felkészítése a zajos körülményekre mindenképpen a gyakorlati használhatóság felé tett lépés, melyet akár az i -vektorok háttérmodelljének (az UBM-nek) a tanítása során is érdemes lenne alkalmazni. (Sajnos itt megint könnyű technikai akadályokba ütközni.)

Tanító adatbázis	Jellemző- kinyerési réteg	Pontosságértékek				
		Pont.	Prec.	Fedés	F_1	AUC
BEA (augmentáció nélkül)	Összegző	58%	60,0%	48,0%	53,3%	0,562
	Szegmens #6	64%	65,2%	60,0%	62,5%	0,628
	Szegmens #7	56%	57,1%	48,0%	52,2%	0,576
BEA (augmentálva)	Összegző	60%	63,2%	48,0%	54,5%	0,595
	Szegmens #6	64%	68,4%	52,0%	59,1%	0,645
	Szegmens #7	58%	61,1%	44,0%	51,2%	0,602
Előtanított modell	Összegző	62%	63,6%	56,0%	59,6%	0,640
	Szegmens #6	70%	72,7%	64,0%	68,1%	0,673
	Szegmens #7	58%	61,1%	44,0%	51,2%	0,527
i-vektor (BEA, augmentáció nélkül)		60%	63,2%	48,0%	54,5%	0,597

2. táblázat. A különböző x-vektorok, valamint a viszonyítási alapként megvizsgált i-vektorok használatával EKZ-azonosításra kapott kiértékelési metrikák. (Pont.: osztályozási pontosság; Prec.: pontosság (precision).)

A három modell közül a legjobb eredményekhez az előtanított modell használata vezetett. Ez véleményünk szerint egyrészt azt támasztja alá, hogy az x-vektorok a gyakorlatban (legalábbis ezen nyelvek esetén) nyelvfüggetlen módon képesek kódolni a beszélőket. Másrészt azt is jelzi, hogy még hatvan órányi felvétel (illetve 165 beszélő) sem képes azt a varianciát nyújtani, amely kellően robusztus x-vektor beágyazások kinyerését lehetővé tevő DNN-ek tanításához szükséges. Kétségtelen, hogy a Snyder és munkatársai által használt korpusz a mintegy hétezer beszélővel lényegesen változatosabb, mint amit akár a teljes BEA adatbázissal (tehát 500 beszélővel) lehetséges lenne elérni.

Összességében elmondható, hogy a kapott pontosságértékek nem különösebben magasak: még a legjobb esethez is csupán 70%-os osztályozási pontosság, és 0,673-es AUC érték tartozik. Ez véleményünk szerint elsősorban a felvételek zajosságának tudható be: a 14 és 35 dB közé eső SNR elég alacsonynak mondható (viszonyításképpen: egy hagyományos analóg telefonvonalhoz 40 dB-es érték tartozik (Aude, 1998)). Ugyanakkor még ezen hátráltató tényező ellenére is jobban el tudtuk különíteni az enyhe kognitív zavarral rendelkező alanyokat az egészséges kontroll személyektől az x-vektorok használatával, mint az i-vektorokra építve.

5. Összegzés

Az enyhe kognitív zavar egy krónikus klinikai szindróma, melynek korai detektálása kulcsfontosságú a kezelés minél hamarabb történő megkezdéséhez. Jelen tanulmányunkban egy viszonylag új jellemzőkinyerési eljárást, az x-vektorokat teszteltük ebben a feladatban. Az x-vektort szolgáltató mély neurális hálókat a

BEA adatbázis egy 60 órás részhalmazán, 165 beszélő adatain tanítottuk két variációban (zaj hozzáadásával és anélkül), valamint egy angol beszédre előtanított modellt is kipróbáltunk. Az x-vektorokat a DNN modellek több rejtett rétegéből is kinyertük.

Eredményeink alapján az x-vektorok valamivel alkalmasabbak az enyhe kognitív zavar detektálására, mint az i-vektorok hasonló méretű adatokon és hasonló (akusztikai) körülmények között. A három tesztelt rejtett réteg közül egyértelműen a mélyebben fekvő szegmensszintű réteg (*Szegmens #6*) vezetett a legjobb eredményekhez mindhárom DNN modell esetében. Az augmentációval tanított modell a legtöbb esetben eredményesebb volt, mint az augmentációs lépés kihagyásával tanított; mindkettő alulmaradt ugyanakkor Snyder és munkatársai előtanított modelljével szemben, melyben valószínűleg a lényegesen nagyobb tanítóadat játszhatott szerepet. Bár kíváncsiak lettünk volna, hogy más keretszintű jellemzők használata hogyan alakítja az eredményeket, a Kaldi beépített x-vektor eszköze meglepően sok korláttal bír: sem a $\Delta / \Delta\Delta$ értékek, sem például frekvenciasávok energiaösszegeinek mint jellemzőknek a használata nem könnyen megoldható. Ugyanígy kíváncsiak lettünk volna rá, hogy a tanítófelvételek „zajosítása” számszerűen mennyit segíthet az i-vektorok teljesítményén, azonban ez az augmentációs lépés is az x-vektor DNN modell tanításához van kötve.

Az osztályozáskor kapott eredményeink nem voltak különösebben átütőek, aminek több oka is lehet. Egyrészt az EKZ-s és kontroll alanyainktól származó felvételek sajnos elég zajosak, melyen utólag már nehéz segíteni (ugyanakkor így talán jobban tükrözik egy valós környezetben lefolytatott EKZ-szűrővizsgálat akusztikai körülményeit). Másrészt érdemes szem előtt tartani, hogy az enyhe kognitív zavart elsősorban a memória és bizonyos nyelvi készségek romlása jellemzi, melyeket sokkal nehezebb kimutatni a beszédből, mint például a Parkinson-kór tüneteit. Mégis, Jeancolas és munkatársai az x-vektorok használatával is „csupán” 70% körüli osztályozási pontosságokat kaptak Parkinson-kór felismerésére (Jeancolas és mtsai, 2020). Természetesen az x-vektor beágyazások lehetséges felhasználása lehet még, hogy kombináljuk azokat (vagy a használatukkal kapott predikciókat) más jellegű jellemzőkkel (például temporális paraméterekkel, lásd Gosztolya és mtsai, 2018), melyet tervezünk kipróbálni a közeljövőben.

Köszönetnyilvánítás

A kutatást részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (projektkód: FK-124413), részben az Innovációs és Technológiai Minisztérium (projektkód: TUDFO/47138-1/2019-ITM) támogatta. Gosztolya Gábor kutatásait az MTA Bolyai János ösztöndíja és az Új Nemzeti Kiválóság Program Bolyai+ pályázata (azonosító: ÚNKP-20-5-SZTE-649) támogatta. A publikációban szereplő kutatást (amelyet a Szegedi Tudományegyetem valósított meg) az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal is támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

Hivatkozások

- Alzheimer’s Association: 2020 Alzheimer’s disease facts and figures. *Alzheimer’s & Dementia* 16(3), 391–460 (2020)
- Aude, A.: Audio quality measurement primer (1998)
- Botelho, C., Teixeira, F., Rolland, T., Abad, A., Trancoso, I.: Pathological speech detection using x-vector embeddings (2020)
- Egas-López, J.V., Tóth, L., Hoffmann, I., Kálmán, J., Pákáski, M., Gosztolya, G.: Assessing Alzheimer’s Disease from speech using the i-vector approach. In: *SPECOM*. pp. 289–298. Isztambul, Törökország (2019)
- García, N., Orozco-Arroyave, J.R., D’Haro, L.F., Dehak, N., Nöth, E.: Evaluation of the neurological state of people with Parkinson’s Disease using i-vectors. In: *Interspeech*. pp. 299–303. Stockholm, Svédország (2017)
- García, N., Vásquez-Correa, J., Orozco-Arroyave, J.R., Nöth, E.: Multimodal i-vectors to detect and evaluate Parkinson’s Disease. pp. 2349–2353. Hyderabad, India (2018)
- Gosztolya, G., Hoffmann, I., Tóth, L., Vincze, V., Pákáski, M., Kálmán, J.: Az enyhe kognitív zavar és korai alzheimer-kór automatikus azonosítása spontán beszédből akusztikus jellemzők segítségével. In: *MSZNY*. pp. 219–230. Szeged (2018)
- de Ipiña, K.L., de Lizarduy, U.M., Calvo, P.M., Beitia, B., García-Melero, J., Fernández, E., Ecay-Torres, M., Faundez-Zanuy, M., Sanz, P.: On the analysis of speech and disfluencies for automatic detection of mild cognitive impairment. *Neural Computing and Applications* 9, 437 (2018)
- Jeancolas, L., Petrovska-Delacrétaz, D., Mangone, G., Benkelfat, B., Corvol, J., Vidailhet, M., Lehericy, S., Benali, H.: X-vectors: New quantitative biomarkers for early Parkinson’s Disease detection from speech. *arXiv preprint arXiv:2007.03599* (2020)
- König, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., Robert, P.H.: Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research* 15(2), 120–129 (2018)
- Meilán, J.J.G., Martínez-Sánchez, F., Martínez-Nicolás, I., Llorente, T.E., Carro, J.: Changes in the rhythm of speech difference between people with nondegenerative mild cognitive impairment and with preclinical dementia. *Behavioural Neurology* 2020, 4683573 (2020)
- Neuberger, T., Gyarmathy, D., Grácsi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: *Proceedings of TSD*. pp. 424–431. Brno, Czech Republic (Sep 2014)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830 (2011)

- Petersen, R.C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V., Fratiglioni, L.: Mild cognitive impairment: a concept in evolution. *Journal of Internal Medicine* 275(3), 214–228 (2014)
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Veselý, K.: The Kaldi speech recognition toolkit. In: *Proceedings of ASRU* (2011)
- Prince, M., Wimo, A., Guerchet, M., Ali, G.C., Wu, Y.T., Prina, M.: *World Alzheimer Report 2015. The Global Impact of Dementia*. Alzheimer’s Disease International, London, UK (2015)
- Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. *Neural Computation* 13(7), 1443–1471 (2001)
- Sluis, R.A., Angus, D., Wiles, J., Back, A., Gibson, T.A., Liddle, J., Worthy, P., Copland, D., Angwin, A.J.: An automated approach to examining pausing in the speech of people with dementia. *American Journal of Alzheimer’s Disease & Other Dementias* 35, 1533317520939773 (2020)
- Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep Neural Network embeddings for text-independent speaker verification. In: *Interspeech*. pp. 999–1003. Stockholm, Svédország (2017)
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust DNN embeddings for speaker verification. In: *ICASSP*. pp. 5329–5333. Calgary, Alberta, Kanada (2018)
- Szatlóczki, G., Hoffmann, I., Vincze, V., Kálmán, J., Pákáski, M.: Speaking in Alzheimer’s Disease, is that an early sign? Importance of changes in language abilities in Alzheimer’s Disease. *Frontiers in Aging Neuroscience* 7, 104943 (2015)
- Themistocleous, C., Eckerström, M., Kokkinakis, D.: Identification of Mild Cognitive Impairment from speech in Swedish using Deep Sequential Neural Networks. *Frontiers in Neurology* 9, 975 (2018)
- Themistocleous, C., Eckerström, M., Kokkinakis, D.: Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *PloS one* 15(7), e0236009 (2020)
- Vincze, V., Szatlóczki, G., Tóth, L., Gosztolya, G., Pákáski, M., Hoffmann, I., Kálmán, J.: Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild Alzheimer’s disease. *Clinical Linguistics & Phonetics* közlésre elfogadva (2020)
- Weiner, J., Schultz, T.: Selecting features for automatic screening for dementia based on speech. In: *SPECOM*. pp. 747–756. Lipcse, Németország (2018)

FORvoice 120+: Statisztikai vizsgálatok és automatikus beszélő verifikációs kísérletek időben eltérő felvételek és különböző beszéd feladatok szerint

Sztahó Dávid, Beke András, Szaszák György

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
1117 Budapest, Magyar tudósok körútja 2.
sztaho.david@vik.bme.hu, beke.andras@gmail.com
szaszak@tmit.bme.hu

Kivonat: A jelen tanulmányban a FORvoice120+ adatbázison végzett akusztikai-fonetikai elemzéseket és automatikus beszélő azonosítási kísérleteket mutatjuk be, a jelenleg elkészült 60 beszélő felvételeivel. Személyfüggő akusztikai jellemzők statisztikai vizsgálatait és automatikus beszélő verifikációs tesztek végzettünk különböző időbeli és beszéd típusbeli eltérések elemzésére. A statisztikai elemzéseknél alaphanghoz, formánsokhoz és beszéd tempóhoz kapcsolódó akusztikai-fonetikai jellemzőket vizsgáltunk. Az eredmények azt mutatták, hogy az eltérő időben történő hangrögzítések alig befolyásolták a jellemzők statisztikai értékeit, míg az eltérő beszédfeladatoknál jelentős eltérés volt tapasztalható. Automatikus beszélő azonosítási (verifikációs) kísérleteket is végeztünk i-vektor és x-vektor implementációkkal. A tesztek alapján elmondható, hogy minél hosszabb beszéd szegmenseket alkalmazunk, annál pontosabb lesz a felismerési eredmény.

1 Bevezetés

Az igazságügyi hangszakértői gyakorlatban az utóbbi időszakban megjelentek és kezdenek elterjedni azok az automatikus módszerek, amelyek egy paradigmaváltás következtében jöttek létre (Morrison, 2011; Saks & Koehler, 2005). Ez a paradigmaváltás igyekszik feloldani azt a kérdést, hogy a mért értékek mennyire tipikusak az egyénre, illetve a populációra nézve. Ez az eljárás mód a kriminalisztika egyéb azonosító technológiáinak módszertanában (pl. DNS azonosítás) is megjelent, és egy egységes összehasonlító rendszert tesz lehetővé, amelybe minen egyéni jellegzetesség mérése beilleszthető valószínűségi értékekkel. Az új paradigma a valószínűségi-arány keretrendszer (likelihood-ratio framework, LR) mennyiségi megvalósítását eredményezi, amely során két hipotézist kell vizsgálni: „Mekkora valószínűséggel származik a kérdéses minta a gyanúsított személytől?”, illetve az ún. ellenhipotézis: „Mekkora valószínűséggel származik a kérdéses minta az adott népességből véletlenszerűen kiválasztott másik személytől?”. Ezek aránya fejezi ki a bizonyítékok erősségét:

$$LR = \frac{p(E|H_{\text{azonos személy}})}{p(E|H_{\text{eltérő személy}})}$$

Az LR rendszerén belül a hang alapú beszélőazonosítási kísérletek elvégzéséhez egy olyan adatbázisra van szükség, amely megfelel az új paradigma alapfeltevéseinek (Beke és mtsai., 2020; Morrison és mtsai., 2012):

- 1) több alkalommal kell minden beszélőtől mintákat rögzíteni (hasonlóság modellezése),
- 2) sok beszélőt kell tartalmaznia lehetőleg a populációra reprezentatíven (a tipikusság modellezéséhez),
- 3) különböző módon rögzített hangmintákat kell felvenni (ún. channel mismatch kompenzálására, pl. telefonos vagy stúdió minőségű),
- 4) egy beszélőtől különböző beszédtypusokat kell rögzíteni a beszédstílus különbségeiből fakadó beszélőn belül is megjelenő eltérések kompenzálására (speech style mismatch compensation).

A jelen cikkben bemutatott kísérletek egy ilyen, most készülő adatbázison valósultak meg. A FORVoice 120+ beszédatadabázis 120 beszélő felvételeit fogja tartalmazni. Ebből jelenleg 60 beszélő felvételei készültek el, amelyeken az eredményeket bemutatjuk. Az adatbázis lehetővé teszi automatikus beszélő azonosítási és -verifikációs kísérletek futtatását, amelyek során eltérő időbeli felvételek és eltérő beszéd feladatok összehasonlítását lehet elvégezni. Ezzel hozzájárul a kriminalisztikai célú hang összehasonlítások módszertanához.

Az automatikus beszélő azonosítás és verifikáció jelenlegi baseline rendszerének, amely illeszkedik a LR-ratio keretrendszerbe, az x-vektorokat használó megvalósítás számít (Snyder és mtsai., 2018). Ez a korábbi i-vektoros megoldást váltotta fel (Dehak és mtsai., 2009, 2010) a deep learning elterjedése által létrejött mély neurális hálózatos megvalósító technikákkal. A felvételeken hallható személyek azonosítása során több eltérő összehasonlítási módot tudunk megkülönböztetni. A beszélők azonosítása (*speaker identification*) során a felismerendő személyazonosság már egy meglévő zárt halmazból kerül ki, tehát előre tudjuk, hogy kik azok a beszélők, akik közül fel kell ismernünk a felvételen hallottat. Ezzel szemben a beszélő verifikáció (*speaker verification*) két beszédminta hasonlóságának mértékét hivatott megállapítani. Ilyen szituációval találkozhatunk tipikusan akkor, amikor egy célszemély azonosságát szeretnénk verifikálni, megerősíteni, hogy tényleg ő hallható a felvételen. Ekkor rendelkezésünkre áll a célszemélytől valamennyi hanganyag, ami alapján egy beszédlenyomatot képzünk, és az igazolni kívánt felvételtől nyert lenyomat ehhez való hasonlóságát szeretnénk mérni.

A jelen tanulmány személyfüggő akusztikai jellemzők statisztikai vizsgálatait és automatikus beszélő verifikációs tesztek eredményeit mutatja be különböző időbeli és beszéd típusbeli eltérések elemzésére. A statisztikai tesztek során megvizsgáltuk, hogy az időben eltérő felvételek és a beszédfeladat típusa befolyásolja-e a mérhető akusztikai-fonetikai paramétereket. Az automatikus beszélő verifikációs kísérletek során pedig az eltérő időtartamú egységek hatását vizsgáltuk. A 2. fejezetben bemutatjuk az adatbázist, utána pedig az elemzéshez alkalmazott eljárásokat írjuk le. Ezután következnek az elért eredmények a 4. fejezetben, majd az ezekből levonható konklúziós és összefoglalás.

2 Adatbázis

A bevezetőben ismertetett igazságügyi hangszakértői kísérletekhez készítendő adatbázis tervezett 120 beszélőjéből 60 beszélő felvétele készült el eddig. A felvételek stúdió minőségű fejmikrofonokkal készültek csendes szobában. A felvételi paraméterek: 44.1kHz mintavételi frekvencia, 16 bites kvantálás, PCM lineáris kódolás. A beszélők (beszélgető partnerek) egy szobában tartózkodtak, egymástól 2-3 méter távolságban. Az adatbázis jelenleg 31 férfi (életkor: 24.2 ± 4.6) és 29 női (életkor: 24.4 ± 5.2) beszélőt tartalmaz. Egy beszélőtől két, eltérő időben készült felvétel került rögzítésre. A két felvétel között két hét telt el minden esetben. Ezeket jelöljük a továbbiakban session 1 és session 2 kifejezésekkel. Minden felvétel három beszédfeladatot tartalmaz (jelölése: task 1-3):

1. Szabad párbeszéd (10 perc).
2. Irányított információcsere (~8 perc): hibás terméklistákon található olvashatatlan információk beszerzése a beszélgető partnertől.
3. A megelőző nap tárgyilagos elmesélése (~3 perc).

A kísérletekhez a felvételeket feldaraboltuk az 500 ms-ot meghaladó időtartamú szünetek mentén. Az 1 másodpercnél rövidebb szakaszokat elhagytuk az így kapott közlések közül. Ezután a közléseket három csoportba osztottuk időtartam szerint: (i) 1-2 mp közöttiek, (ii) 2-5 mp közöttiek, valamint (iii) 5 mp felettiek.

3 Eljárások

3.1 Statisztikai tesztek

A munka során különböző statisztikai teszteket végeztünk, amelyekben különböző akusztikai jellemzők eltérését vizsgáltuk bizonyos szempontok szerint. Az alkalmazott statisztikai eljárás a Generalized Linear Mixed Models (GLMM) volt (McCulloch & Neuhaus, 2014). A teszteket az SPSS 22.0 (Corp, 2013) verziójával valósítottuk meg. A GLMM során az alkalmazott *id*-k a beszélők azonosítói voltak, a *fixed effects* változók pedig a *session* és *task* azonosítók. Ilyen módon mérhetőek voltak, hogy a felvétel időpontja és a feladat típusa szerint kimutatható-e statisztikai eltérés a mért akusztika-fonetikai értékek között. Ez információt ad arról, hogy mennyire alkalmazhatók az időben, illetve a feladat típusok szerint eltérő felvételek a személyek azonosítása során.

3.2 Akusztikai-fonetikai jellemzők

Az akusztikai-fonetikai jellemzőket a Praat (Boersma, 2001) segítségével számítottuk ki. A következő jellemzőket vizsgáltuk (a későbbiekben használt jelölésüket zárójelben tüntettük fel).

(1) A felvételenként számított *alaphang értékek átlaga és szórása* ($f0.avg$, $f0.std$). A számítási ablakméret 50 ms volt, 10 ms-os lépésközzel. Minden felvételre kiszámítottuk az összes alaphang értéket (ahol zöngés hangok fordultak elő), és ezekenek vettük az átlagát és szórását.

(2) *Artikulációs sebesség* (art_tempo). Minden felvételre kiszámítottuk az artikulációs sebességet.

(3) *Pairwise variability indices* ($rPVIc$, $rPVIV$, $nPVIc$, $nPVIV$). Felvételenként kiszámítottuk külön a magánhangzó és mássalhangzó időtartamok időtartamának változásának mérőszámát (Grabe & Low, 2002). Két változatot alkalmaztunk: nyers (raw) és normalizált, amelyeket a következő módon számítottunk:

$$rPVI = \left[\sum_{k=1}^{m-1} \frac{|d_k - d_{k+1}|}{(m-1)} \right] \text{ és}$$

$$nPVI = 100 * \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}} \right| / (m-1) \right],$$

ahol a d_k a k . fonémát jelöli, az m pedig a fonémák teljes számát. Mindkét jellemzőt külön kiszámítottuk a mássalhangzókra és a magánhangzókra is (jelölésük: $rPVIc$, $rPVIV$, $nPVIc$, $nPVIV$).

(4) Felvételenként az /e/ és /a/ hangokon számolt *első három formáns és sáv szélességük átlaga és szórása* ($E.fl.avg$, $E.flbw.avg$, $E.fl.std$, $E.flbw.std$, a további formánsok és az /a/ hang hasonlóan jelölve). A számításnál 25 ms-os ablakméretet és 10 ms-os lépésközt alkalmaztunk.

A jellemzőket beszédfeladatonként és *session*-önként számoltuk ki, tehát pl. egy alaphang átlagértéket számoltunk a *session 1* és *task 1* felvételen, egyet a *session 1 task 2* felvételen, és így tovább. A jellemzők kiszámításánál nem vettük figyelembe a szünetek mentén való darabolást.

3.3 Beszélő verifikáció

A statisztikai vizsgálatok után automatikus beszélő verifikációs kísérleteket végeztünk az eddig elkészült adatbázison annak érdekében, hogy megvizsgáljuk, hogy az eltérő időtartamú felvételek mennyire befolyásolják a gépi modellekkel kapott eredményeket.

Az automatikus beszélő verifikációs kísérletekhez *i*-vektor (Dehak és mtsai., 2009) és *x*-vektor (Snyder és mtsai., 2018) alapú megoldásokat használtunk fel. A megvalósítások KALDI keretrendszerben készültek David Snyder receptjei alapján, amelyek jelenleg state-of-the-art *baseline* megoldásoknak számítanak (*i-vector and x-vector KALDI recipe*, 2018).

Az *i*-vektor implementálása során 512 keverékszámú GMM-UBM modellt használtunk, az *i*-vektorok mérete pedig 100 volt. Az *x*-vektor esetén a tanított TDNN 400 dimenziós volt. Akusztikai jellemzőként mindkét megoldásnál 12 MFCC-t alkalmaztunk. Szintén mindkét eljárás esetén PLDA-val (Ioffe, 2006) történt a tesztesetek kiértékelése. Az eredményeket *equal error rate* (EER) szerint mértük.

A felvételeket két halmazra osztottuk. A 60 beszélő közül 40-et használtunk a i-vektor kinyerő, az x-vektor TDNN valamint a PLDA tanítására. A maradék 20 beszélő

1. Táblázat: Az akusztikai-fonetikai jellemzőkre kapott, GLMM-el mért p értékek a *task* és *session* változókra. A 99%-on szignifikáns** és 95%-on szignifikáns* eltéréseket külön jelöltük.

jellemző	<i>task</i>	<i>session</i>
art_tempo	0.000**	0.068
rPVlc	0.000**	0.105
rPVlv	0.000**	0.185
nPVlc	0.000**	0.153
nPVlv	0.042*	0.286
E.f1.avg	0.023*	0.769
E.f2.avg	0.000**	0.213
E.f3.avg	0.000**	0.578
E.f1.std	0.000**	0.864
E.f2.std	0.000**	0.225
E.f3.std	0.000**	0.402
E.f1bw.avg	0.169	0.698
E.f2bw.avg	0.001**	0.715
E.f3bw.avg	0.534	0.171
E.f1bw.std	0.103	0.484
E.f2bw.std	0.021*	0.654
E.f3bw.std	0.605	0.274
O.f1.avg	0.103	0.971
O.f2.avg	0.000**	0.970
O.f3.avg	0.177	0.700
O.f1.std	0.000**	0.620
O.f2.std	0.000**	0.582
O.f3.std	0.001**	0.786
O.f1bw.avg	0.000**	0.593
O.f2bw.avg	0.032*	0.564
O.f3bw.avg	0.048*	0.825
O.f1bw.std	0.160	0.788
O.f2bw.std	0.086	0.172
O.f3bw.std	0.674	0.200
F0 avg	0.001**	0.579
F0 std	0.814	0.032*

felvételeit alkalmaztuk tesztelésre a következő módon. A *session 1*-be tartozó felvételek kerültek az *enrollment* halmazba, vagyis ezek voltak azok a felvételek, amelyeken a beszélők átlagvektorait számítottuk ki. A *session 2*-be tartozó felvételek voltak a konkrét tesztesetek (*target*), amelyekben a személyazonosságot meg kellett állapítani.

Az előzetes kísérletek azt mutatták, hogy a 40 beszélő hanganyaga nem elegendő a TDNN és a PLDA tanítására, ezért ehhez még felhasználtuk a BABEL (Roach és mtsai., 1996) és az MRBA (Vicsi & Vig, 1998) adatbázisokat is, amelyekben összesen 388 beszélő szerepelt, összesen 120 percnyi beszéddidővel.

4 Eredmények

4.1 GLMM eredmények

A GLMM-el végzett elemzések statisztikai eredményeit az 1. táblázat tartalmazza. Minden jellemzőhöz megadtuk a feladat (*task*) és felvételi időpont (*session*) szerinti p értéket. A 95%-os, valamint a 99%-os szignifikancia szintű eltéréseket külön kiemeltük (* és ** jelölések). A jellemzőket a 3.2 fejezetben leírt jelölésekkel láttuk el.

A *session* változó esetén csupán egyetlen esetben tértek el a mért értékek statisztikailag egymástól, az alaphang szórásakor. Minden más esetben azt mondhatjuk, hogy a felvételek időben eltérő rögzítése nem volt hatással arra, hogy a mért értékek jelentősen eltérnek-e egymástól.

A beszédfeladatok esetén pont az ellenkező jelenséget tapasztaltuk. Csupán néhány olyan jellemző van, amelynél nincs szignifikáns különbség a *task* változó értékeinek függvényében.

A beszéd tempóra vonatkozó jellemzők (artikulációs tempó és PVI jellemzők) mind szignifikáns eltéréseket mutatnak az eltérő beszédfeladatoknál, míg az időbeli eltérés nem volt számottevő hatással a mért értékekre. A formánsok esetén vegyes szignifikanciájú eltéréseket mértünk a beszédfeladatok között, ám az időben eltérő felvételek itt sem mutattak sehol sem eltéréseket. Az átlagos alaphang értékek szignifikánsan megváltoztak, ha eltérő beszédfeladról volt szó, ám nem voltak eltérő populációból tekinthetőnek, ha az időbeli eltéréseket nézzük. Ezzel szemben az alaphangok felvételenkénti szórása pont a *session* változó szerint volt eltérő, míg a *task* változó nem volt rá jelentős hatással. Az összes eredményt tekintetbe véve azt mondhatjuk, hogy az időben eltérő felvételek kevésbé (alig) befolyásolják a beszélőkre jellemző értékeket, az eltérő beszédfeladatok viszont jelentős hatással vannak, ezért azokat majd érdemes figyelembe venni a későbbiekben a beszélő azonosítási kísérletek során.

4.2 Automatikus beszélő verifikáció

Az automatikus beszélő verifikációs kísérletek során azt vizsgáltuk meg, hogy a különböző időtartamú beszédszakaszok hogyan befolyásolják az beszélők azonosításának pontosságát. Ehhez a következő teszteseteket végeztük el:

- (a) minden beszédszakasz felhasználása (*all*),

- (b) a 1-2 mp közötti időtartamú beszédszakaszok külön alkalmazása (tanítás és tesztelés) ($1-2v1-2$),
- (c) a 2-5 mp közötti időtartamú beszédszakaszok külön alkalmazása (tanítás és tesztelés) ($2-5v2-5$),
- (d) az 5 mp feletti időtartamú beszédszakaszok külön alkalmazása (tanítás és tesztelés) ($5v5$),
- (e) 1-2 mp időtartamú felvételekkel történő tanítás, és 5 mp feletti időtartamú felvételekkel való tesztelés ($1-2v5$),
- (f) 5 mp feletti időtartamú felvételekkel való tanítás, 1-2 mp időtartamú felvételekkel való tesztelés ($5v1-2$). A kapott eredményeket a 2. táblázat tartalmazza, ahol a teszteseteket az előzőekben leírt módon jelöltük. A táblázatban a PLDA pontozással kapott EER% értékeket tüntettük fel. Mivel jelenleg még nincs elegendő hanganyag a beszédfeladatok külön alkalmazására, ezért az összes beszédfeladatot felhasználtuk a kísérletek során.

Az eredmények azt mutatják, hogy az i-vektoros megvalósítás alacsonyabb tévesztési százalékokat produkál annak ellenére, hogy az x-vektoros rendszer elvileg újabb technológiának számít. Ez azért lehet, mert az x-vektor kinyerés mély tanuláson alapul, így tanításukhoz sokkal több hanganyag szükséges, mint az i-vektorhoz. A jelenlegi adatbázis nem éri el az a méretet, amivel a TDNN háló tanítható (az MRBA és BABEL kiegészítéssel együtt sem). A nemzetközi irodalomban sem egyértelmű az x-vektor alapú megközelítés felsőbbrendű helyzete (Kanagasundaram és mtsai., 2011; Sarkar és mtsai., 2012).

Az összes hanganyaggal elvégzett kísérletek (*a* eset) eredménye (5.4% EER) összehasonlítható a nemzetközi irodalommal (Snyder és mtsai., 2018). Ahogy azt várni lehetett, az 5 mp-nél hosszabb felvételekkel kaptuk a legjobb eredményt (a hosszabb minták jobban leírják a beszélőt). Ennek megfelelően a legrövidebb minták (1-2 mp) adták a legrosszabb azonosítást (7.727% a 3.193%-hoz képest).

2. Táblázat: Az automatikus beszélő verifikáció eredményei. PLDA-val kapott EER %-ok az i-vektor és x-vektor implementációkra.

Teszteset	i-vektor	x-vektor
<i>all</i>	5.405	9.276
<i>1-2v1-2</i>	6.605	11.38
<i>2-5v2-5</i>	3.957	6.345
<i>5v5</i>	3.193	1.739
<i>1-2v5</i>	3.193	1.91
<i>5v1-2</i>	7.727	10.56

5 Konklúzió

Az akusztikai-fonetikai paraméterek elemzése alapján elmondható, hogy a felvételek időbeli eltérése nem mutatott jelentős eltérést a mérésekben. Tehát ez a változó nem

okoz zavart akkor, amikor beszélő verifikációt, azonosítást végzünk. Csupán az, hogy egy adott személytől különböző időben rögzítünk hanganyagot, nem befolyásolja az azonosítást (ha egyéb beszédképzést befolyásoló tényező, például megfázás, nem jelentkezik).

Ezzel ellentétben, a beszéd stílusát meghatározó változó (jelen esetben a beszéd feladat) jelentős hatással volt a mérhető eltérésekre. A monológok és a szabad párbeszéd során megfigyelhető volt olyan eltérés, amely szignifikánsnak mutatkozott. Célszerű tehát egy adott személytől sokféle beszéd helyzetet rögzíteni, ha személyazonosítást megvalósító feladatról van szó.

Az automatikus beszélő azonosítást célzó kísérletek során az megnyilatkozások időtartama (*utterances*) hatással vannak a beszélő azonosítás pontosságára. Minél hosszabb felvétel áll rendelkezésre, annál jobb eredményt lehet elérni az általánosan elterjedt i-vektor és x-vektor alapú rendszerrel. 5 másodpercnél hosszabb felvételek esetén 1.739% EER-t lehet elérni.

A jelenleg rendelkezésre álló hanganyag 60 beszélőt tartalmaz. A végső tervezett 120 beszélővel már robosztusabb eredményeket és elemzéseket lehet majd elkészíteni. Ezen kívül ez már elegendő lesz ahhoz is, hogy a beszéd feladatok közötti eltéréseket automatikus verifikációs kísérletekkel vizsgáljuk.

6 Összefoglalás

A jelen tanulmányban a FORvoice120+ adatbázison végzett akusztikai-fonetikai elemzéseket és automatikus beszélő azonosítási kísérleteket mutattuk be, a jelenleg elkészült 60 beszélő felvételeivel.

A statisztikai mérésekhez alaphangból, formánsokból és beszéd tempóhoz kapcsolódó akusztikai-fonetikai jellemzőket alkalmaztuk. Az eredmények azt mutatták, hogy az eltérő időben történő hangrögzítések alig befolyásolták a jellemzők statisztikai értékeit, míg az eltérő beszédfeladatoknál jelentős eltérés volt tapasztalható.

Automatikus beszélő azonosítási (verifikációs) kísérleteket is végeztünk i-vektor és x-vektor implementációkkal. A tesztek alapján elmondható, hogy minél hosszabb beszéd szegmenseket alkalmazunk, annál pontosabb lesz a felismerési eredmény.

Köszönetnyilvánítás

Az FK128615 számú projekt a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással, az FK pályázati program finanszírozásában valósult meg.

Hivatkozások

Beke, A., Szaszák, G., & Sztahó, D. (2020). FORvoice 120+: Magyar nyelvű utánkövetéses adatbázis kriminalisztikai célú hangösszehasonlításra. In G. Berend, G. Gosztolya, & V. Vincze

- (Szerk.), *XVI. Magyar Számítógépes Nyelvészeti Konferencia* (o. 95–101). Szegedi Tudományegyetem, Informatikai Intézet; MTMT. <https://m2.mtmt.hu/api/publication/31148107>
- Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott. Int.*, 5(9), 341–345.
- Corp, I. B. M. (2013). IBM SPSS statistics for windows, version 22.0. *Armonk, NY: IBM Corp.*
- Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., & Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. *Tenth Annual conference of the international speech communication association*.
- Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), 788–798.
- Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, 7(515–546).
- Ioffe, S. (2006). Probabilistic linear discriminant analysis. *European Conference on Computer Vision*, 531–542.
- I-vector and x-vector KALDI recipe*. (2018). <https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16>
- Kanagasundaram, A., Vogt, R., Dean, D. B., Sridharan, S., & Mason, M. W. (2011). I-vector based speaker recognition on short utterances. *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2341–2344.
- McCulloch, C. E., & Neuhaus, J. M. (2014). Generalized linear mixed models. *Wiley StatsRef: Statistics Reference Online*.
- Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51(3), 91–98.
- Morrison, G. S., Rose, P., & Zhang, C. (2012). Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. *Australian Journal of Forensic Sciences*, 44(2), 155–167.
- Roach, P., Arnfield, S., Barry, W., Baltova, J., Boldea, M., Fourcin, A., Gonet, W., Gubrynowicz, R., Hallum, E., Lamel, L., Marasek, K., Marchal, A., Meister, E., & Vicsi, K. (1996). BABEL: An Eastern European multi-language database. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 3, 1892–1893 köt.3. <https://doi.org/10.1109/ICSLP.1996.608002>
- Saks, M. J., & Koehler, J. J. (2005). The coming paradigm shift in forensic identification science. *Science*, 309(5736), 892–895.
- Sarkar, A. K., Matrouf, D., Bousquet, P. M., & Bonastre, J.-F. (2012). Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. *Thirteenth Annual Conference of the International Speech Communication Association*.
- Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333.
- Vicsi, K., & Vig, A. (1998). First Hungarian speech database. *Beszédkutatás*, 98, 163–177.

SPONTÁN BESZÉD, CHAT

A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata

Vincze Veronika¹, Üveges István^{2,3}, Szabó Martina Katalin^{3,4}, Takács Károly^{4,5}

¹MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.

²Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
6722 Szeged, Egyetem utca 2.

³Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.

⁴Társadalomtudományi Kutatóközpont, CSS-RECENS
1097 Budapest, Tóth Kálmán utca 4.

⁵Linköpingi Egyetem, The Institute for Analytical Sociology
601 74 Norrköping, Svédország
{vinczev,martina}@inf.u-szeged.hu
uvegesistvan898@gmail.com
Szabo.Martina@tk.hu
karoly.takacs@liu.se

Kivonat A tanulmányban egy nagyméretű, magyar, beszélt nyelvi adatbázist elemzünk, és annak morfológiai és szófaji sajátosságait vetjük össze más írott nyelvi korpuszok sajátosságaival. A HuTongue korpusz, amelyet manuálisan leiratoztattunk és annotáltattunk, elsősorban abból a célból készült, hogy egy alapvetően szociológusokból álló kutatócsoport a pletykadiskurzusok sajátosságait vizsgálhassa (Galántai és mtsai, 2018). A korpusz szövegei hétköznapi situációkban, külső ingerektől elszigetelt környezetben keletkeztek (Gulyás és mtsai, 2018). Legjobb tudomásunk szerint a HuTongue az első olyan, nagyméretű, magyar beszélt nyelvi korpusz, amely szüretlenül tartalmazza az összes, az adott időszakban elhangzó rögzített beszélgetés részletesen annotált leiratát. Mivel az egyes beszélgetések esetenként több résztvevő oldalán is rögzítésre kerültek, így a duplikátumok kezelése a feldolgozás egy fontos, nem triviális lépése volt. A jelen részletesen tárgyalja e munkafázis megfontolásait és módszereit. Ezt követően bemutatja a létrejövő korpusz statisztikai, köztük morfológiai és szófaji alapadatait, összevetve néhány más írott korpusz alapvető adataival. Azt reméljük, hogy a korpuszunk hatékonyan támogatja majd számos különféle szociológiai és nyelvészeti probléma korpuszalapú kutatását a jövőben.

Kulcsszavak: kézzel annotált korpusz, nyelvi erőforrás, írott és beszélt nyelv, pletyka, magyar, NLP

1. Bevezetés

Manapság egyre több olyan kutatási terület van, közöttük a szociológia vagy a nyelvészet, amely mindinkább adatközpontúvá válik. Ezek a korpuszalapú és statisztikai megközelítések azonban megbízható és nagyméretű nyelvi adatbázisok létrehozását teszik szükségessé (Neuberger és mtsai, 2014). Közülük egyre több törekszik a beszélt nyelv reprezentálására (Crowdy, 1993; Hemphill és mtsai, 1990; Maekawa és mtsai, 2000; Oostdijk, 2000; Mengusoglu és Deroo, 2001; Seppänen és mtsai, 2003; Van Bael és mtsai, 2007).

A tanulmányban bemutatunk egy nagyméretű, magyar beszélt nyelvi adatbázist, amelyet manuálisan leiratoztattunk és annotáltattunk. Az adatbázis kifejezetten a pletykadiskurzusok sajátosságainak vizsgálatához készült, azonban a korpusz mérete és a szövegek sajátosságai miatt számos egyéb kutatási kérdés tárgyalásához is alapot teremthet a jövőben.

A korpusz létrehozásának fő célja a pletyka fogalmi körébe sorolható megnyilatkozások egzakt vizsgálata volt, elkészítése pedig három fő szakaszra tagolódott. A feldolgozást egy előkészítési fázis előzte meg, amelyet követően a fájlokat legépelték és annotálta egy feldolgozócsoporthoz. E munka során tehát az annotátorok nem csupán legépelték a hanganyagok verbális tartalmát, hanem kódolták a nem verbális hanghatásokat, valamint a pletykadiskurzusokat és az utóbbiak célszemélyeit is. Az utolsó fő lépésként a kutatócsoportnak ki kellett szűrnie azokat a duplikátumokat, amelyek a felvételőrgyítési sajátosságok miatt kerültek a korpuszba.

A jelen dolgozat kettős céllal bír: Egyrészt a cikk részletesen tárgyalja az utolsó munkafázis megfontolásait és módszereit. Másrészt bemutatja a létrejövő korpusz statisztikai, köztük morfológiai és szófaji alapadatait, összevetve néhány más írott korpusz alapvető adataival. Célunk, hogy felmérjük, milyen jellegzetes eltéréseket tapasztalhatunk a szófaji eloszlás és morfológiai jellegzetességek tekintetében az írott és a beszélt nyelv között. Azt reméljük, hogy a korpuszunk hatékonyan támogatja majd számos különféle szociológiai és nyelvészeti probléma korpuszalapú kutatását a jövőben, valamint az írott és a beszélt nyelv összevető vizsgálataihoz is adalékot szolgáltat.

2. Kapcsolódó irodalom

A korpuszok növekvő száma ellenére még mindig viszonylagosan kevés a hangzó szövegeket reprezentáló száma, különösen azoké, amelyek gépelt leiratokat is tartalmaznak. Ez az átírási eljárás magas munkaerő- és költségigényével magyarázható. Különösen csekély a magyar beszélt nyelvű beszélt korpuszok száma, és ezek is többségükben olvasott szövegekből állnak (Gósy és mtsai, 2012). Az alábbiakban csupán egy összefoglalást adunk a magyar nyelvű beszélt korpuszairól, mindezek részletesebb bemutatását l. (Szabó és mtsai, 2021).

A magyar telefonbeszéd adatbázis (MTBA) telefonon rögzített olvasott szövegeket tartalmaz. Feldolgozási módját úgy alakították ki, hogy támogathassa a beszédtechnológiai kutatásokat és fejlesztéseket (Vicsi és mtsai, 2002). A Kivi

korpusz (Kugler, 2015) különféle történetek elmeséléseiből áll, míg a Budapesti Szociolingvisztikai Interjú 250 adatközlő interjút tartalmazza (Várad, 2003). A HuComTech multimodális korpusz körülbelül 50 órnyi video- és hangfelvételtől, összesen 111 formális (szimulált állásinterjú) és 111 informális, de irányított párbeszédből áll (Pápay és mtsai, 2011).

A fentebbiektől eltérően a spontán beszédet kívánja reprezentálni a Budapesti Egyetemi Kollégiumi Korpusz (BEKK) (Bodó és mtsai, 2017) és a BEszélt nyelvi Adatbázis (BEA) (Gósy, 2013), és ezzel összefüggésben a HuTongue szövegállománya az említettek közül a két utóbbihoz áll a legközelebb. (A kutatócsoport egy újabb, hasonló korpuszról l. (Szabó és mtsai, 2021).) A BEKK esetében az interakciókat a résztvevők saját telefonjaikon rögzítették, ezért tulajdonképpen társalgásrészleteket tartalmaz. A BEA korpusz létrehozóinak fő célja az volt, hogy fonetikai (és nem szemantikai vagy pragmatikai) vizsgálatokat tegyen lehetővé, ezért a korpuszban alkalmazott annotációt is ennek megfelelően alakították ki. A szövegek létrejöttének körülményei, illetve feldolgozásuk módja miatt azonban a fentebb említett korpuszok csupán korlátozottan alkalmasak a magyar beszélt nyelv sajátosságainak a kutatására.

Legjobb tudomásunk szerint a HuTongue az első olyan, nagyméretű, magyar beszélt nyelvi korpusz, amely tartalmazza az összes beszélgetés részletesen annotált leiratát, ezáltal képes hatékonyan támogatni számos nyelvészeti, valamint szociológiai tárgyú kutatást, valamint, mivel egy zárt közösség adott időintervallumban elhangzott valamennyi beszélgetését tartalmazza, kiemelten alkalmas lehet kvantitatív nyelvészeti elemzésekre is (Szabó és Galántai, 2017).

3. A korpusz létrehozásának menete

Ebben a részben összefoglaló jelleggel ismertetjük a korpusz létrehozásának metódusát és eszközeit (részletesen Szabó és Galántai (2017); Gulyás és mtsai (2018); Galántai és mtsai (2018); Pápay (2019)). A folyamat fő részei a következők voltak:

- előfeldolgozás: eltávolítottuk a hosszabb csendeket és felosztottuk az anyagot kisebb egységekre,
- zajok kiszűrése Python függvénykönyvtárakkal,
- a hanganyag leírása, annotálása és a fájlok adatbázisba rendezése,
- minőségbiztosítás a teljes folyamat során,
- a duplikátumok eltávolítása,
- automatikus morfológiai elemzés a magyarlanc programcsomaggal (Zsibrita és mtsai, 2013),
- kvantitatív mérések kivitelezése a korpusz szűrt verzióján.

3.1. A korpusz anyaga és az előfeldolgozási lépések

A korpusz szövegei hétköznapi kommunikációs helyzetekben keletkeztek, egy külső ingerektől elzárt környezetben (Galántai és mtsai, 2018). A magas minőségű

hanganyagot egy szórakoztatóipari cég rögzítette. A hang rögzítését 24 órában végezték, a keletkezett korpusz pedig összesen 8 egymást követő nap felvételeit tartalmazza. A felvételek készítése során mind a nyolc önkéntes résztvevő mikroportot viselt; beszélgetéseik teljes rögzítéséhez előzetes beleegyezésüket adták.

A szövegek abban a tekintetben spontán beszélgetések, hogy azok témáit és hosszát a felvételek készítői nem határozták meg, továbbá a beszéd mennyiségét sem szabályozták. A résztvevők tehát korlátozás nélkül beszéltek annyit és arról, amennyit és amiről akartak, ugyanakkor tisztában voltak azzal a ténnyel, hogy a hangjukat folyamatosan rögzítik. Emellett néhány esetben a résztvevők cselekedeteit külső irányítással befolyásolták. Mivel ily módon a szövegek keletkezési körülményei (a résztvevők motivációi, valamint az alkalmankénti külső irányítás) befolyásolhatták a beszélői megnyilatkozásokat, a korpuszt félig vagy részlegesen spontánnak kell tekintenünk (Szabó és Szvetelszky, 2019).

A korpusz előkészítésének első lépéseként a 10 másodpercnél hosszabb csendeket automatikusan eltávolítottuk (ezáltal szegmenseket képeztünk, l. lentebb), és az anyagot 60 perces egységekre osztottuk fel. A létrehozott hangzószöveg-korpusz körülbelül 500 órányi anyagot tesz ki.

3.2. Hanganyagok leiratozása és annotálása

A korpusz építésének második lépéseként 18 annotátor legépelte és annotálta a fájlokat. A feldolgozási munkát a megfelelő résztvevők kiválasztása és képzése előzte meg (részletesen l. Szabó és Galántai (2017)). A munka során az annotátorok az f4transcript szoftvert¹ használták. Mind a leiratozást, mind az annotálást ennek a segítségével végezték. Az f4 szoftvert gyakran használják szociológiai tárgyú tudományos kutatási projektekben, mivel kiváló lehetőséget ad nagyobb mennyiségű hanganyag gyors és egy időben történő leiratozására és tagelésére. Mivel projektünk célja nem az volt, hogy fonetikai elemzésekhez állítsunk elő vizsgálati anyagot, a választott szoftver megfelelő volt a számunkra.

A gépelés és annotálás színvonalának biztosítása érdekében a munka során gyakran ellenőriztük a minőséget úgy, hogy bizonyos fájlokat az összes annotátorral feldolgoztattuk, majd a kimeneteket összevetettük egymással (részletesen l. Gulyás és mtsai (2018)). A leiratok pontossága érdekében az annotátorokkal rendszeres megbeszéléseken tisztáztuk az esetleges inkonzisztenciákat, a leiratozást végzők személye pedig néhány esetben cserélődött is a minőségi elvárások tarthatósága érdekében.

Az annotátorok online kapták meg a hangfájlokat, valamint az egyes audiofájlokhoz tartozó szegmenshatárokat tartalmazó egyszerűszöveg-formátumú fájlokat. Az annotátorok tehát ezeket a fájlokat töltötték be a szoftverbe, és a szövegeket a megfelelő szegmenshatárok közé gépelték a kapott utasításoknak megfelelően.

Mivel minden résztvevő viselt mikroportot és minden mikroport anyagát feldolgoztuk, nem volt szükséges az anyagok teljes tartalmát legépelni. Az alapelv az volt, hogy leírjuk azt a beszélgetést, amelyben a mikroport viselője részt vesz.

¹ <https://www.audiotranskription.de/english/f4>

Az annotátorok feladata a következő három részfeladatból állt (Gulyás és mtsai, 2018):

- a hanganyagon rögzített verbális kommunikáció legépelése,
- az anyag vizsgálata szempontjából fontos, különféle információk kódolása (időbélyegek, az adott diskurzusok résztvevői, valamint a beszélgetések során jelen levő, de meg nem szólaló résztvevők),
- különféle nem verbális hanghatások tagelése az annotálási útmutatóban előre meghatározott módon (pl. suttogás, kiabálás, nevetés, sóhaj stb.),
- a pletykadiskurzusok megjelölése a pletyka célszemélyével / személyeivel egyetemben.

Ahogy az annotációs folyamat fentebb részletezett lépései is mutatják, a munka célja nem kizárólag a verbális tartalmak leírása volt, hanem bizonyos, nonverbális információk annotálása is. Ezzel összefüggésben olyan nem verbális jeleket választottunk ki az annotáláshoz, amelyeknek gyakorisága vagy együttes előfordulási jellemzői a pletyka indikátoraként szolgálhatnak (Galántai és mtsai, 2018).

A gépelőknek időbélyegek segítségével el kellett különíteniük a beszélgetések egyes megnyilatkozásait, illetve összefüggő monológjait. Ezt azt jelentette, hogy egyazon időbélyeg alá kerülhetett egyetlen megnyilatkozás vagy egy összefüggő monológ, de több résztvevő megszólalását külön időbélyegek alá kellett tenni. Az adott megszólaló nevét is jelölték. Az időbélyegek nagyon fontosak voltak a számkra, mivel ezek a címkék nyújtanak lehetőséget a különböző mikroportokon rögzített, ezáltal különálló anyagokon létező szegmensek egymáshoz illesztésére. Mindemellett, ezek a tagek tartják meg a kapcsolatot az audiófájlok és azok írott változatai között.

Azoknak a megnyilatkozásait, akik nem voltak a vizsgálati csoport tagjai, egy speciális annotációs címkével látták el. Emellett azokat a személyeket is annotálták, akik nem szólaltak meg egy adott beszélgetés során, azonban jelen voltak (taggel jelölték a nevüket, vagy ha nem voltak azonosíthatóak, legalább a becsült számukat).

Azt is annotálni kellett, ha egy megnyilatkozás vagy egy beszélgetés egésze vagy egy része érthetetlen volt. Ezen túlmenően, ha az annotátor nem volt biztos abban, hogy jól értette az elhangzottakat, bizonytalanságát egy speciális nyitó- és zárótaggal jelölte. A nem verbális hangok (pl. köhögés, nevetés) két alapvető típusát különbözőképpen kódoltattuk: a pillanatnyit és a hosszabb ideig tartót.

Az annotációs folyamat kardinális lépése volt azoknak a megszólalásoknak a címkézése, amelyben a résztvevők valamely más, jelen nem levő résztvevőre utalnak. Ebben a vizsgálatban elsődlegesen a csoporton belüli pletykára fókuszálunk, így Kurland és Pelled (2000) alapján a pletykát kutatócsoportunk a következőképpen határozta meg: megnyilatkozás vagy beszélgetés valamely csoport általában néhány tagja között az adott csoport más olyan tagjáról vagy tagjairól, aki vagy akik nincs(nek) jelen. Amennyiben a pletyka célszemélye az annotátor számára egyértelmű volt, akkor ezt egy megfelelő annotációs címkével ugyancsak fel kellett tüntetnie.

Ahogy az a munkafolyamat vázlatából is kitűnik, az annotálás a leiratozással, így a felvett anyag hallgatásával egy időben zajlott, tehát nem utólag végeztettük a munkát a gépelt anyagon. Az annotátoroktól azt kértük, hogy az annotálás során a hangsúlyt és a hanglejtést éppúgy vegyék figyelembe, és azok segítségével próbálják megérteni a szó szerinti jelentésen túli, szándékolt tartalmakat is, valamint azokra támaszkodva hozzanak döntést a kétes esetekben.

3.3. A duplikátumok eltávolítása

Mivel minden résztvevő mikroportot viselt, néhány beszélgetést többször is rögzítettek a felvételeken. A feldolgozást nehezítette az a körülmény, hogy ezen rögzített „beszélgetéspéldányok” száma nem egyezett meg az abban részt vevők tényleges számával. Az eltérés számos faktorból adódhatott, mint például:

- az egyik résztvevő felvételén bizonyos részletek túlságosan halkak voltak / nem álltak rendelkezésre a pontos leírat elkészítéséhez,
- túlságosan erős háttérzaj (ilyen esetekben egyáltalán nem készülhetett leírat),
- esetenként leiratozói hanyagság miatt.

A felvételek résztvevőnként eltérő hossza és szegmentálása miatt nem volt egyértelmű továbbá, ha egy adott beszélgetést már korábban rögzítettek valahol a korpuszban. Az annotátorok ezért azt az utasítást kapták, hogy írjanak le minden elhangzottat, függetlenül attól, hogy az adott beszélgetést már esetlegesen hallották egy másik mikroporton keletkezett anyag leiratozásakor, ugyanakkor feltéve, hogy az adott beszélgetésben a mikroport viselője megítélésük szerint részt vesz, valamint a beszélgetés felismerhető minőségben szerepel az adott felvételen (pl. nem túl távoli vagy zajos).

Ezeknek a duplikátumoknak a kiszűrése nyilvánvalóan kardinális feladat, hiszen enélkül a kvantitatív eredmények bármely kutatási kérdés vonatkozásában szignifikánsan eltérhetnek az ismétlődésmentes változat eredményeitől. Annak céljából tehát, hogy a vizsgálataink előtt a korpuszból a duplikátumokat eltávolíthassuk, a következő eljárást alkalmaztuk.² A fájlokat a 8 napon történő rögzítés okán 8 csoportba soroltuk: mindegyik csoport az adott napon rögzített felvételekből állt. Ezután összegyűjtöttük az ugyanazon a napon rögzített összes beszélgetés szókincsét a szegmenshatárokkal egymástól elválasztott diskurzusonként, azaz minden beszélgetéshez készítettünk egy szógyakorisági listát (bag-of-words). Ezután összehasonlítottuk az egyes beszélgetések szókincsét az összes többi beszélgetés szókincsével, amelyeket a többi résztvevő mikroportjai rögzítettek ugyanazon a napon. Ha két diskurzus között a szókincs legalább 75%-a egyezett, és az adott beszélgetések legalább 10 szót tartalmaztak, akkor az adott két beszélgetést azonosnak tekintettük³. Az esetek többségében egy hosszabb beszélgetés tartalmazott egy rövidebb szekvenciát, azaz a rövidebb szekvencia

² Az volt a célunk, hogy minél egyszerűbb és hatékonyabb megoldást válasszunk.

³ Több küszöbértékkel is kísérleteztünk, és 75% bizonyult a leghatékonyabbnak.

megismétlődött a korpuszban. Ezeknek az ismétléseknek az eltávolítása érdekében a rövidebb beszélgetést töröltük az adatokból.⁴

A fenti megközelítés hatékonyságát manuálisan kiértékeljük az adatok egy kis részhalmazán; az ellenőrzéshez 50 diskurzuspárt ellenőriztünk manuálisan. Megállapítottuk, hogy 45 esetben (90%) a törölt beszélgetést egy másik, hosszabb beszélgetés valóban tartalmazta. Azt mondhatjuk tehát, hogy megközelítésünk képes volt 90%-os pontossággal (precision) megtisztítani a korpuszt az eredményeket torzító duplikátumoktól.

Módszerünkkel a három vagy annál többször előforduló szövegrészeket is lehetséges volt eltávolítanunk. A páronkénti összehasonlítás során ugyanis értelem-szerűen minden diskurzust minden diskurzussal összevetettünk, így többszörös ismétlődések esetében is csupán a meghatározottak szerint legmegfelelőbb példányt tartottuk meg.

A munka során a kiinduló adatbázis 35,7%-át töröltük (15470-ből 5519 szegmens), így a HuTongue fennmaradó része (1 469 558 token) már alkalmasabb lehet más, már létező korpuszokkal való összevetésre.

A duplikátumok eltávolításának a bemutatott megoldáson túl több alternatívája is lehetséges. Az egyik lehetőség az, ha az annotációkat súlyozzuk a diskurzusban résztvevők számával. Ez azt jelenti, hogy amennyiben egy beszélgetésben például öten vettek részt, úgy az adott beszélgetés leiratában annotált információkat 1/5 részben számítjuk bele a statisztikai adatokba. A lehetséges további alternatívákkal, azok alkalmazhatóságával azonban e dolgozat keretei között nem foglalkozunk.

4. A szűrt korpusz alapvető statisztikai adatai

Ebben a fejezetben a HuTongue szűrt változatának alapvető statisztikai adatait mutatjuk be.

4.1. Annotált elemek

A korpusz annotálása során összesen 78 486 taget helyeztünk el. A nem verbális hangok statisztikai alapadatait az 1. táblázat tartalmazza.

Látható, hogy összességében több mint 50 000 hangeffektus található a rögzített anyagban, azaz a spontán beszéd számos ponton tartalmaz nem verbális elemeket. Ezek egy része a szórakoztatóipari műsor jellegéből adódik (pl. ujjongás, fütyülés, sikítás), míg más hangeffektusok előfordulhatnak nem spontán jellegű rögzített beszédben is (pl. köhögés egy hírműsor felvételében). Az automatikus beszédfelismeréshez azonban ezen elemekre mindenképpen érdemes figyelmet fordítani.

A pletykaannotáció statisztikai alapadatait a 2. táblázat tartalmazza. A nevet adatvédelmi okok miatt lecseréltük.

⁴ Mivel az időbélyegek relatív, és nem abszolút időhatárokat jelöltek, nem volt lehetőség egyszerű időbélyegalapú szűrésre.

Típus	Előfordulás
köhögés	936
sóhajítás	3333
nevetés	32777
sírás	505
gunyoros nevetés	1420
zavarodott nevetés	1870
sikítás	579
ásítás	293
pisszegés	28
ujjongás	875
torokköszörülés	2118
fütyülés	596
éneklés	6441
összesen	51771

1. táblázat. Nem verbális hangok eloszlása.

Az adatokból egyértelműen látszik, hogy bizonyos személyeket (Zoli, Maja, András) jelentősen többször említene, mint másokat – nevezetesen, a tagek közel 48%-a róluk szólt. Valószínűleg ők állnak a társaság életének középpontjában, több és erősebb kapcsolati hálóval rendelkeznek, mint a kevésbé gyakran emlegetett személyek. Ezzel szemben Zsani, Zsáklin, Viola és Dóri a csoport marginálisabb tagjának számítanak, ők kevésbé képezték a dialógusok központi témáját. A csoportközi említéshálózat mélyebb elemzésével egy másik dolgozatban foglalkozunk (Üveges és mtsai, 2021).

Célszemély	Előfordulás
Zoli	4430
Maja	3775
András	3502
Gabi	2406
Dani	1930
Vanda	1760
Kornél	1689
Sanyi	1199
Tomi	1157
Erika	925
Levi	605
Zsani	472
Zsáklin	421
Viola	219
Dóri	160
összesen	24650

2. táblázat. A pletykaszövegek eloszlása célszemély szerint.

4.2. Szófaji eloszlás

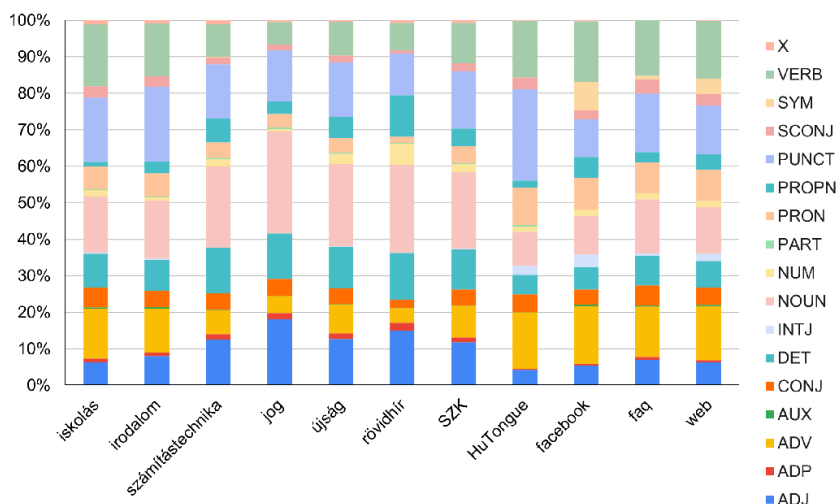
A HuTongue korpusz szövegeit a magyarlanc nyelvi elemzővel (Zsibrita és mtsai, 2013) elemeztük, majd az így kapott szófaji eloszlást összevetettük több kézzel annotált írásos szöveg szófaji eloszlásával. Ezzel azokra a kérdésekre kerestük a választ, hogy milyen jellegzetes eltéréseket tapasztalhatunk szófaji eloszlás terén az írott és beszélt nyelv között. Vizsgálatainkhoz a HuTongue korpusz mellett a Szeged Korpusz univerzális morfológiára (Vincze és mtsai, 2017) annotált változatára, valamint két kisebb, közösségi médiából származó adatbázisra (Vincze és mtsai, 2014) támaszkodtunk. E két utóbbi korpusz a Facebookról gyűjtött nyilvános posztokat, valamint a www.gyakorikerdesek.hu oldalról származó kérdéseket és válaszokat tartalmaz, és szintén az univerzális morfológia szerint lettek kézzel annotálva.

A fent említett korpuszok szófaji statisztikáit a 3. táblázat, valamint az 1. ábra tartalmazza. A korpuszok hasonlóságát az úgynevezett Kendall-együtthatóval számszerűsítettük, lásd a 4. táblázat.

Szófaj	iskolás	irodalom	sz.tech.	jog	újság	rövidhír	SzK	HuTongue	FB	FAQ	Web
ADJ	21267	18641	26496	46190	27799	33698	174091	61701	467	681	1148
ADP	3304	2510	2934	4269	3597	4936	21550	4411	38	78	116
ADV	46592	28201	14099	12006	17275	9093	127266	227503	1369	1332	2701
AUX	797	807	153	57	225	79	2118	562	29	33	62
CONJ	18816	10587	9551	11856	9553	5141	65504	71403	363	532	895
DET	31253	19793	26160	31495	25196	29027	162924	79276	530	785	1315
INTJ	738	814	114	6	135	5	1812	36514	288	60	348
NOUN	52385	37299	47041	71445	49590	54445	312205	136819	921	1441	2362
NUM	6102	2040	4125	1695	6246	13128	33336	21077	151	152	303
PART	956	884	709	1077	642	505	4773	7108	0	0	0
PRON	21227	14654	9188	9585	9001	3646	67301	149479	746	834	1580
PROP	3901	7702	13807	8638	12553	25861	72462	27778	487	271	758
PUNCT	59420	47990	31241	35820	32902	25755	233128	367820	904	1551	2455
SCONJ	10521	6761	3978	4001	4142	1994	31397	48155	202	366	568
SYM	0	0	350	1	1	59	411	33	670	102	772
VERB	57905	33998	18805	15500	20526	16834	163568	227023	1440	1477	2917
X	3496	1930	2222	1386	794	1633	11461	2896	19	0	19

3. táblázat. A HuTongue, a Szeged Korpusz egyes doménjeinek, valamint egyes közösségi média-szövegek szófaji megoszlása.

Az eredmények azt mutatják, hogy több szembevetendő szófaji gyakorisági különbséget is tapasztalhatunk az írott és beszélt nyelv között. A Szeged Korpusz egészét tekintve az írott nyelvben átlagosan jelentősen több melléknév és főnév fordul elő, míg a beszélt nyelvben az igék, határozószók, indulatszavak és névmások szerepe nő meg. Érdekes ugyanakkor megfigyelni, hogy a Szeged Korpusz egyes doménjei is eltérően viselkednek e téren. A Kendall-együtthatót is figyelembe véve a HuTongue szövegeihez a gyakori kérdések, az irodalmi, valamint az iskolás alkörpuszok állnak a legközelebb. A legnagyobb távolságot pedig a szófaji eloszlás terén az üzleti rövidhírek mutatják.



1. ábra: A szófajok eloszlása.

A kommunikatív célok alapján a vizsgált korpuszokat két nagyobb csoportra oszthatjuk:

- Interaktív korpuszok: a HuTongue mellett ide sorolhatjuk az iskolás, irodalmi, gyakori kérdések korpuszokat is. Elsődleges jellemzőjük, hogy a szerzőnek / beszélőnek határozott szándéka, hogy megszólítsa, illetve párbeszédet folytasson az olvasóval / beszélgetőpartnerrel. A HuTongue és a gyakori kérdések esetében ez a párbeszéd forma magától értetődik, ugyanakkor az irodalmi szövegekben (regényekben) is számos párbeszéd rész található. Az iskolás szövegek létrehozásakor a tanulók pedig azt az instrukciót kapták, hogy meséljenek a hallgatóságnak egy számukra kedves napról, illetve érveljenek egy téma mellett. Mindkét szövegfajtában számos, a közönség felé szóló „kiszólást” találunk a korpuszban. E domének hasonlóságát a Kendall-együttható is alátámasztja.
- Leíró korpuszok: a jogi szövegek, újsághírek, üzleti hírek és számítástechnikai szövegek fő célja az olvasó tényszerű informálása, azonban az interakció szerepe itt jóval kisebb, az olvasó szerepe majdnem kizárólagosan az információ befogadására korlátozódik. Meg kell említenünk ugyanakkor, hogy az újságok és a számítástechnikai magazinok interjúkat is tartalmaznak, melyek a párbeszéd forma miatt közelebb állnak az interaktív korpuszokhoz, így e két domén némileg közelebb áll az interaktív szövegekhez, ahogy a Kendall-együttható is mutatja.

A Facebookról származó szövegek ugyancsak változatosak az interakció szempontjából: egyrészt különféle márkákat, sztárokat stb. képviselő oldalak nyilvános informatív bejegyzései kerültek ide (kommentek nélkül), másrészt személyes jellegű (de nyilvános láthatóságú) bejegyzéseket is találunk itt. A kétfajta bejegyzés

célja megint csak eltérő, így nem meglepő, hogy a Facebook-szövegek is valahol középen helyezkednek el az együttthatósági skálán.

Az interaktív szövegek szófaji jellemzői tehát az alábbiakban foglalhatók össze. Gyakoriak bennük az igék (az olvasót / hallgatót cselekvésre buzdítják) és a névmások (élőbeszédben vagy az ahhoz közel álló írott szövegekben gyakoriak a deiktikus utalások). Az indulatszavak szerepe is kiemelkedő, ezek nyomatékosítják adott esetben a mondanivalót, máskor figyelemfelhívó szereppel bírnak stb. A határozószavak gyakorisága pedig az igék gyakoriságával függhet össze: igék mellett határozószavak jelennek meg, ellenben főnevek mellett melléknevek tudják kifejezni ugyanazt a minőséget. Míg a leíró jellegű korpuszokban a főnevek és melléknevek szerepe domináns, addig ugyanazt a jelentéstartalmat a dinamikusabb ige + határozószó pár fejezi ki az interaktív korpuszokban.

Korpuszrész	Hasonlóság
iskolás	0,9534
irodalom	0,9559
számítástechnika	0,8909
jog	0,9007
újság	0,9007
rövidhír	0,7978
Szeged Korpusz	0,9081
Facebook	0,8848
Gyakori kérdések	0,9558
Webes szövegek	0,9363

4. táblázat. A HuTongue hasonlósága a Szeged Korpusz egyes doménjeihez, valamint egyes közösségimédia-szövegekhez a szófaji eloszlás alapján.

5. Összegzés

A dolgozatban bemutattuk és kontrasztív módon elemeztük a HuTongue korpusz újabb, duplikátumoktól megtisztított változatát. A HuTongue a magyar beszélt nyelvet reprezentálja, a hangzó szövegek legépelt és annotált változatával együtt.

A tanulmány célja az volt, hogy összefoglalja a korpuszkészítés fő lépéseit és módszereit, majd ismertesse azt a megoldást, amellyel sikeresen kiszűrtek a korpuszban található többször előforduló azonos szövegrészeket. A korpusz a szöveganyaga, mérete, valamint a szövegek feldolgozási módja miatt lehetőséget ad számos olyan kutatás elvégzésére, amely a magyar beszélt nyelv valamely sajátosságát veszi górcső alá. Mindemellett a duplikátumok kiszűrésével a korpuszelemzés kvantitatív eredményei is mentesülnek a torzító tényezők alól.

Második fő lépésként bemutattuk a korpusz újabb változatának alapvető adatait, és azokat összevetettük néhány más szövegkorpusz morfológiai és szófaji gyakorisági sajátosságaival. Az összevetés során rámutattunk néhány olyan hasonlóságra és eltérésre, amely az egyes korpuszok, illetve szövegtípusok és -domének

között mutatkozik. Bízunk benne, hogy a HuTongue korpusz további érdekes adalékokkal járulhat hozzá a magyar spontán beszélt nyelv különböző vizsgálataihoz.

Tervezzük a korpusz nyilvánossá tételét a jövőben a kutatók számára, az érzékeny adatok anonimizálását követően.

Köszönetnyilvánítás

A kutatást az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében támogatta (ERC_CoG_2014_648693 sz. szerződésben), a kutatás vezetője Takács Károly.

Szabó Martina Katalin kutatásait részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal – NKFIH OTKA posztdoktori kiválósági programja (NKFI-azonosító: 132312) támogatta.

Hivatkozások

- Bodó, Cs., Kocsis, Zs., Vargha, F.: A Budapesti Egyetemi Kollégiumi Korpusz. Elméleti és módszertani kérdések. In: Benő, A., Fazakas, N. (szerk.) Élőnyelvi kutatások és a dialektológia: Válogatás a 19. Élőnyelvi Konferencia - Marosvásárhely, 2016. szeptember 7-9. - előadásaiból. pp. 169–177 (2017)
- Crowdy, S.: Spoken corpus design. *Literary and Linguistic Computing* 8(4), 259–265 (1993)
- Galántai, J., Pápay, B., Kubik, B.G., Szabó, M.K., Takács, K.: A pletyka a társas rend szolgálatában – az informális kommunikáció struktúrájának mélyebb megértéséért a computational social science eszközeivel. *Magyar Tudomány* 179(7), 964–976 (2018)
- Gósy, M.: BEA–A multifunctional Hungarian spoken language database. *Phonetician* 105, 50–61 (2013)
- Gósy, M., Gyarmathy, D., Horváth, V., Grácsi, T.E., Beke, A., Neuberger, T., Nikléczy, P.: Bea: Beszélt nyelvi adatbázis (2012)
- Gulyás, A., Galántai, J., Szabó, M.K., Szebeni, Z.: A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai. In: MSZNY 2018 - XIV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 317–330 (2018)
- Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The ATIS spoken language systems pilot corpus. In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania* (1990)
- Kugler, N.: Megfigyelés és következtetés a nyelvi tevékenységben. Tinta Könyvkiadó (2015)
- Kurland, N.B., Pelled, L.H.: Passing the word: Toward a model of gossip and power in the workplace. *Academy of management review* 25(2), 428–438 (2000)
- Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous Speech Corpus of Japanese. In: *LREC*. pp. 947–9520. Citeseer (2000)

- Mengusoglu, E., Deroo, O.: Turkish LVCSR: Database Preparation and Language Modeling for an Agglutinative Language. In: IEEE International Conference on Acoustics Speech And Signal Processing. vol. 6, pp. 4018–4018. IEEE (2001)
- Neuberger, T., Gyarmathy, D., Grácsi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: International Conference on Text, Speech, and Dialogue. pp. 424–431. Springer (2014)
- Oostdijk, N.: The Spoken Dutch Corpus. Overview and First Evaluation. In: LREC. pp. 887–894. Athens, Greece (2000)
- Pápay, B.: The Purpose and Types of Organizational Gossip. Ph.D.-értékezés (2019)
- Pápay, K., Szeghalmy, Sz., Szekrényes, I.: Hucomtech multimodal corpus annotation. *Argumentum* 7, 330–347 (2011)
- Seppänen, T., Toivanen, J., Väyrynen, E.: MediaTeam speech corpus: a first large Finnish emotional speech database. In: Proceedings of the Proceedings of XV International Conference of Phonetic Science. pp. 2469–2472. Citeseer (2003)
- Szabó, M.K., Galántai, J.: Egy magyar nyelvű spontán beszélt nyelvi korpusz (HuTongue) létrehozásának tapasztalatai. In: XXVI. MANYE Kongresszus konferenciakötete. Pécs (2017)
- Szabó, M.K., Szvetelszky, Zs.: Részlegesen spontán körülmények között keletkezett pletykaszövegek pragmatikai szempontú vizsgálata. *Nyelvtudományi Közlemények* 115, 317–343 (2019)
- Szabó, M.K., Vincze, V., Ring, O., Üveges, I., Vit, E., Samu, F., Gulyás, A., Galántai, J., Szvetelszky, Zs., Bodor-Eranus, E.H., Takács, K.: StaffTalk: magyar nyelvű spontán beszélgetések korpusza. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)
- Üveges, I., Szabó, M.K., Vincze, V.: Szó, beszéd – avagy hogyan kommunikálunk egymásról. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)
- Van Bael, C., Baayen, R.H., Strik, H.: Segment deletion in spontaneous speech: a corpus study using mixed effects models with crossed random effects. In: INTERSPEECH. pp. 2741–2744 (2007)
- Váradi, T.: A budapesti szociolingvisztikai interjú. In: Kiefer, F., Siptár, P. (szerk.) *A magyar nyelv kézikönyve*. pp. 339–359. Akadémiai Könyvkiadó, Budapest (2003)
- Vicsi, K., Tóth, L., Kocsor, A., Csirik, J.: MTBA—a Hungarian telephone speech database. *Híradástechnika*, LVII 8 (2002)
- Vincze, V., Simkó, K.I., Szántó, Zs., Farkas, R.: Universal Dependencies and morphology for Hungarian - and on the price of universality. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 356–365. Association for Computational Linguistics, Valencia, Spain (Apr 2017), <https://www.aclweb.org/anthology/E17-1034>

- Vincze, V., Simkó, K.I., Varga, V.: Annotating uncertainty in Hungarian web-text. In: Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop. pp. 64–69. Association for Computational Linguistics and Dublin City University, Dublin, Ireland (Aug 2014), <https://www.aclweb.org/anthology/W14-4909>
- Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. pp. 763–771 (2013)

Magyar nyelvű spontán beszéd szemantikai–pragmatikai sajátosságainak elemzése nagy méretű korpusz (StaffTalk) alapján

Vincze Veronika¹, Üveges István^{2,3}, Szabó Martina Katalin^{2,4}

¹MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.

²Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.

³Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
6722 Szeged, Egyetem utca 2.

⁴Társadalomtudományi Kutatóközpont, CSS-RECENS
1097 Budapest, Tóth Kálmán utca 4.

{vinczev,martina}@inf.u-szeged.hu, uvegesistvan898@gmail.com

Kivonat A dolgozatban bizonyos pragmatikai és szemantikai sajátosságokat vizsgálunk magyar nyelvű, nagy méretű spontánbeszéd-korpusz (StaffTalk) alapján. A vizsgálati korpusz is hétköznapi szituációkban, külső hatásoknak is kitett munkahelyi környezetben, spontán módon létrejött nyelvi tartalmakból áll, vagyis a kutatásban résztvevők szabadon megválaszthatták beszélgetésük tárgyát, hosszát és partnereit. A korpusz létrehozása során a hanganyagokat legépelték, majd azt követően számos szempont alapján annotálták. A jelen vizsgálatokat ezeket az annotációkat felhasználva végezzük el.

Kulcsszavak: korpusz, spontán beszéd, magyar, pragmatika, beszédaktusok, udvariasság, bizonytalanság

1. Bevezetés

A dolgozatban a magyar spontán beszéd bizonyos szemantikai és pragmatikai sajátosságait vizsgáljuk nagy méretű, kézzel annotált korpusz alapján. A cikk hiánypótló, hiszen a magyar spontán beszéd bizonyos szemantikai és pragmatikai sajátosságait vizsgálja egy egyedülálló, nagy méretű, kézzel annotált spontánbeszéd-korpusz alapján. A hiány oka alapvetően az, hogy még nemzetközi szinten is szerény azoknak az adatbázisoknak a száma, amelyek a spontán beszédet reprezentálják, illetve azoké is, amelyek valamilyen kézzel készített, szemantikai–pragmatikai annotációval rendelkeznek.

A vizsgálati korpusz, amely az első olyan, magyar nyelvű spontánbeszéd-adatbázis, amely számos szemantikai és pragmatikai sajátosság kézi annotációját tartalmazza, hétköznapi szituációkban, spontán módon létrejött nyelvi tartalmakból áll, amelyek külső hatásoknak is kitett munkahelyi környezetben

keletkeztek. A beszélgetések rögzítése egy magyarországi iskola épületében zajlott 27 munkanapon keresztül. A tanári munkaközösség azon tagjai (összesen 20 fő), akik önként vállalták a kutatásban való részvételt, egy okosórát viseltek, mellyel felvették egymás közti beszélgetéseiket. A hangfájlokat egy annotátor-csapattal legépeltettük és három különböző fázisban annotáltattuk, amelyek a következők voltak: pletykadiskurzusok, bizonyos pragmatikai sajátságok, valamint bizonytalanságra utaló nyelvi elemek.

A jelen dolgozatban tett megállapításokhoz a két utóbbi szinten létrehozott annotációt használjuk fel. A közösségen belüli kommunikáció vizsgálatának egyik fontos vetülete, hogy milyen beszédaktusokat és udvariassági stratégiákat használnak egymás között az egyes közösségi tagok. Ugyanakkor a nyelvi bizonytalanság külön is figyelmet érdemel és kifejezőeszközei sokszor egybeesnek bizonyos beszédaktusokkal. Mindezek beható vizsgálatára, valamint összefüggéseinek feltárására ad lehetőséget a StaffTalk korpusz részletes annotációja. Az elvégzett korpuszvizsgálatok új eredményeket hozhatnak a spontán beszélt nyelv diskurzusaktusainak kvantitatív és kvalitatív sajátságairól, amelyek összevethetően a korábbi, elméleti szintű nyelvészeti megállapításokkal.

2. Kapcsolódó irodalom

Bár a nemzetközi irodalomban egyre több spontánbeszéd-adatbázissal találkozni (Crowdy, 1993; Hemphill és mtsai, 1990; Maekawa és mtsai, 2000; Oostdijk, 2000; Van Bael és mtsai, 2007; Neuberger és mtsai, 2014), a magyar nyelvű korpuszok száma ezen a téren messze elmarad a nemzetközitől. Ugyanakkor mind a nemzetközi, mind a hazai spontánbeszéd-vizsgálatokra egyaránt jellemző, hogy azok alapvetően fonetikai, illetve akusztikus sajátságok elemzésére irányulnak (pl. Kane és mtsai (2011); Reichel és Mády (2013); Deme és Markó (2013); Lenne és mtsai (2009); Zhu és Penn (2006)).

A nemzetközi korpuszok közül a legtöbb, amely pragmatikai annotációt is tartalmaz, írott szövegekből (De Felice és mtsai, 2013) vagy, amennyiben hangzó szövegeket tartalmaz – telefonbeszélgetésekből készült (Leech és mtsai, 2003). Így például a brit Telecom 1200 telefonbeszélgetéséből készült OASIS korpusz beszédaktus-szintű annotációt tartalmaz (Leech és mtsai, 2003). A Switchboard korpusz, amelyet több különböző sajátság mentén is annotáltak, szintén tartalmazza a beszédaktusok tagjeit is (Calhoun és mtsai, 2010). A dialógusaktusok jelentősen több típusát annotálták a fentebbi Switchboard korpusz egy részén. A munka célja az volt, hogy vizsgálati és tanító anyagot készítsenek a természetes-nyelvi interakció statisztikai modellezésére és a diskurzusstruktúrák automatikus detektálására (Jurafsky és mtsai, 1997). Más, nem telefonbeszélgetéseket tartalmazó spontánbeszéd-korpuszok nemzetközi szinten is kifejezetten ritkák (pl. (Cheng és mtsai, 2005)).

Ami a magyar nyelvet illeti, jelenleg egyetlen olyan magyar, beszélt nyelvi korpuszról van tudomásunk (HuComTech), amely diskurzus szintű annotációt is tartalmaz, azonban ez az annotáció mindössze négy sajátságra terjed ki (turn-taking, turn-giving, backchannel, turn-keeping) (Pápay és mtsai, 2011).

A kutatócsoportunk által készített másik nagyméretű korpusz, a HuTongue (Galántai és mtsai, 2018; Gulyás és mtsai, 2018) csupán félig (vagy részlegesen) spontánnak tekinthető, mivel egy szórakoztató jellegű tévéműsor céljaira készültek a felvételek, és, bár a társalgások a legtöbbször nem voltak kívülről kérdésekkel vagy témameghatározásokkal irányítva, a szövegek keletkezési körülményei (a résztvevők motivációi, valamint az időnkénti rendezői irányítás) befolyásolhatták a beszélői megnyilatkozásokat. Ugyanakkor olyan szemantikai–pragmatikai sajátságok, mint például bizonyos beszédaktusok, a nyelvi udvariasság különböző formái vagy a bizonytalanság beható vizsgálata csupán nagy méretű, kézzel megfelelően annotált spontánbeszéd-korpusz alapján lehetséges.

A fejezet további részében néhány sajátság kiemelésével szeretnénk jobban rámutatni az elkészített korpusz annotációjának a fontosságára.

Danescu-Niculescu-Mizil és mtsai (2013) alapján az udvariasság például a humán kommunikáció központi motorja, éppoly alapvető, mint az igazmondás, az informativitás, a relevancia vagy a világosság követelményei (Paul és mtsai, 1975; Leech, 2016; Brown és Levinson, 1978). A természetes nyelvben az udvariasság számtalan eszközzel és variációban kódolható (Danescu-Niculescu-Mizil és mtsai, 2013). Markerei szorosan kapcsolódnak a társadalmi interakciók hatalmi dinamikájához, és gyakran meghatározó tényezők abban, hogy ezek az interakciók jól vagy rosszul működnek-e (Andersson és Pearson, 1999; Rogers és Lee-Wong, 2003; Holmes és Stubbe, 2015).

A beszédaktusok kutatásának egyik központi kérdése a különbségtétel a lokúciós és az illokúciós jelentés között (Austin, 1975; De Felice és mtsai, 2013). Röviden összefoglalva, az előbbi a beszédaktus szó szerinti, kimondott vagy leírt tartalmára utal (pl. túl alacsony vagyok ahhoz, hogy elérjem a polcot), míg utóbbi a beszédaktus funkcióra utal, ami a beszélő a megnyilatkozásával valójában kommunikálni szándékozik (pl. segítség kérése a magas polcon lévő tárgy eléréséhez). Az, hogy hogyan lehet helytállóan elszámolni ezzel a különbséggel, a beszédaktus annotációjának egyik fő kihívása (De Felice és mtsai, 2013).

Az ún. indirekt beszédaktusok esetében, Searle (1975) definíciója alapján egy adott beszédaktust a beszélő egy másik beszédaktussal valószínűsít meg. E beszédaktusok megfelelő kezelésének kiemelt jelentősége van többek között a mesterséges intelligencia területén, hiszen ennek hiánya csökkenti az intelligens rendszerek hatékonyságát az emberi környezettel való interakcióban (Roque és mtsai, 2020).

3. Az annotált korpusz

A StaffTalk korpusz hétköznapi szituációkban, spontán módon létrejött nyelvi tartalmakból áll, amelyek külső hatásoknak is kitett munkahelyi környezetben keletkeztek 27 munkanap alatt. A korpuszt spontán nyelvi produktumok alkotják, vagyis a kutatásban résztvevők szabadon megválaszthatták beszélgetésük tárgyát, hosszát és partnereit. A résztvevők okosórát viseltek, melyek rögzítették beszélgetéseiket. (Mindezekről részletesebben beszámolunk egy másik, ugyanezen a konferencián megjelent dolgozatban (Szabó és mtsai, 2021)).

Az órák összesen 215:26:18 időtartamú hanganyagot rögzítettek. A projekt előkészítő szakaszában, első lépésként a hangfájlokból kivágtuk a tíz másodpercnél hosszabb csendeket, majd az anyagot tovább válogattuk: kiszűrtük a kutatás szempontjából nem releváns, adatvédelmi szempontból problémás, valamint nagyon rossz minőségű fájlokat. Az előválogatás után 105:16:10 időtartamú hanganyag maradt (közel 47%-a az eredeti felvételeknek), a feldolgozás során ennek leiratozása, majd annotálása történt meg.

A leiratozási fázisban tíz gépelő vett részt, akik a hallott anyagot legépeltek, időbélyegekkal, illetve különféle annotációkkal látták el. Elengedhetetlen volt az egyes diskurzusokban részt vevő személyek név szerinti azonosítása is. (Mindezekről ugyancsak részletesen beszámol Szabó és mtsai (2021).) E fázist követően a létrejött szövegfájlokat három különálló fázisban annotáltattuk, amelyhez az MMAX2 eszközt (Müller és Strube, 2006) használtuk. A munka során a pletykadiskurzusokat, különböző pragmatikai sajátságokat, valamint a nyelvi bizonytalanság jelölőit annotáltattuk sokrétűen.

Ebben a fejezetben részletesen bemutatjuk a korpusz két, pletykán kívüli annotációját, valamint közlünk néhány megállapítást az annotáció alapján végzett statisztikai vizsgálatokról.

3.1. Pragmatika

A különböző pragmatikai jelenségek esetében – amennyiben lehetséges volt – (minimum) teljes tagmondatokat jelöltünk. Ha egymás után több tagmondat/mondat is ugyanabba a kategóriába tartozott (pl. hosszasan panaszkodott valaki), akkor azt egy egységként jelöltük be.

A pragmatikai egységek típusát illetően egyaránt figyelembe vettük az Austin-Searle neve által fémjelzett beszédaktuselméletet (Austin, 1975; Searle, 1975), udvariasságelméleteket (Brown és Levinson, 1978), valamint az ezekre adott lehetséges reakciókat, valamint az irónia és antiirónia jelenségeit. Különálló kategóriaként vettük fel a „figyelem felhívása” beszédaktust, mivel úgy véljük, hogy spontán beszélgetésekben ennek kiemelt szerepe és gyakorisága lehet, a beszélő-partnerek személyes interakciójának köszönhetően. Hangsúlyoznunk kell azt is, hogy több, hagyományosan különállónak tekintett beszédaktust összevontunk a jelen annotációs sémában, elsősorban azért, mert pusztán a leírt és hallott beszédre támaszkodva nem kaphatunk teljes képet a beszélő motivációjáról, szándékairól, érzelmeinek erősségéről, ami például a figyelmeztetés és fenyegetés elkülönítésében kulcsfontosságú szerepet kapna.

A nyelvi bizonytalanság annotálásakor azt a minimális egységet/kulcsszót (szót vagy szókapcsolatot) jelöltük, amely önmagában is felelős volt a bizonytalanságért, pl. *talán*, *lehet*, *szerintem*, *nem is tudom* stb. Ható és feltételes módú igék esetében, amennyiben bizonytalan jelentéstartalommal rendelkeztek, a teljes igét jelöltük (mivel morfémat nem tudtuk önmagában kijelölni). Több szó együtt tehát kizárólag akkor volt jelölhető, ha együtt hordozta a bizonytalan tartalmat (pl. *tudom* vs. *nem tudom*).

Ami a nyelvi bizonytalanság típusait illeti, sok esetben valamely lexikális tartalom, másképpen egy konkrét nyelvi elem felelős a bizonytalanságért egyfaj-

ta bizonytalansági markerként. Más típusú bizonytalanságok esetében azonban nem lehet csupán a szemantikára koncentrálni, ugyanis a bizonytalanságot a ko-, illetve kontextus határozza meg. Az előbbi a fentebbieknek megfelelően szemantikai, az utóbbit diskurzusszintű bizonytalanságnak nevezzük, és azoknak több altípusát különböztetjük meg (Vincze, 2013).

Mindkét annotációs szint esetében azt kértük az annotátoroktól, hogy a munkát a hanganyag hallgatásával egyszerre végezzék, és a jelöléseket mindig az aktuális kontextus és hangsúly, illetve hanglejtés függvényében végezzék el.

A beszédaktusokat és pragmatikai sajátságokat az alábbi annotációs séma szerint annotáltuk. (A kevésbé egyértelműekhez rövid magyarázatot fűzünk.)

- Beszédaktusok:
 - ígéret / ajánlat (jövőbeli pozitív cselekedetre utalás)
 - figyelmeztetés / fenyegetés (jövőbeli negatív cselekedetre utalás)
 - kérés / parancs / kívánság
 - panasz / vád / kritika / sértés (a partner vagy bármely személy (a beszélő maga is lehet) iránti negatív vélemény kifejezése negatív jelentéstartalmú szavakkal)
 - dicséret / bók (a partner vagy bármely személy (a beszélő maga is lehet) iránti pozitív vélemény kifejezése pozitív jelentéstartalmú szavakkal)
 - bocsánatkérés
 - köszönetnyilvánítás
- Reakciók:
 - elfogadás / egyetértés
 - visszautasítás / egyet nem értés (nyílt visszautasítás/egyet nem értés; ajánlatra, kérésre adott direkt visszautasítás vagy az egyet nem értés nyílt kifejezése)
 - hárítás (ajánlatra, kérésre adott válaszként, nem derül ki, hogy egyetért vagy nem ért egyet az előzőekkel, tehát ez az egyet nem értés vagy visszautasítás „kikerülése”)
- Irónia:
 - irónia (a beszélői szándék szerint a szótári jelentéssel ellentétes értékkel használt szavakkal kifejezett megnyilatkozások, tehát pozitív jelentéstartalmú szavakkal kifejezett negatív tartalom)
 - antiirónia (negatív jelentéstartalmú szavakkal kifejezett pozitív értékelés)
- Interakciós elemek:
 - figyelem felhívása (fontos vagy érdekes mondandó jelzése a partner felé)
 - üdvözlés / elköszönés

A pragmatikai annotációt képzett nyelvészek végezték az MMAX2 szoftver (Müller és Strube, 2006) segítségével.

3.2. Bizonytalanság

A bizonytalanság annotálásakor követtük a már korábban létrehozott magyar nyelvű bizonytalansági korpuszok kategorizálását (Vincze, 2014, 2016), melyet az alábbiakban foglalunk össze:

- Szemantikus bizonytalanság:
 - episztemikus: a világtudásunk alapján nem tudjuk eldönteni, hogy igaz-e vagy hamis az állítás. Gyakran ható igékkel fejeződik ki, de más lexikai elemek is előfordulhatnak (*talán, valószínűleg, lehetséges*)
 - doxasztikus: hiedelmek, vélemény kifejezése (*hisz, gondol, vél, szerint*)
 - feltételes: egy adott feltételhez kötött az állítás igazságértéke (*ha... akkor*)
 - vizsgálat: pl. kutatási kérdés egy tudományos cikkben (*megvizsgál, elemez*)
- Diskurzus szintű bizonytalanság:
 - weasel: bizonytalan információforrás vagy szereplő a cselekvésben (*valaki, egyesek*)
 - hedge: mennyiségek vagy minőségek homályos jelölése (*sok, gyakori*)
 - peacock: bizonyít(hat)atlan állítás vagy túlzás (*gyönyörűség, botrányos*)

A bizonytalanság annotálását – a pragmatikaihoz hasonlóan – képzett nyelvészek végezték az MMAX2 szoftver (Müller és Strube, 2006) segítségével.

4. Eredmények

Ebben a fejezetben összegezzük a kétféle annotációs szint eredményeit, valamint röviden elemezzük a pragmatikai és bizonytalansági annotáció kapcsolatát.

4.1. Pragmatikai annotáció

Az annotált korpuszban található pragmatikai annotált elemek gyakoriságát az 1. táblázat szemlélteti.

Az adatokból kiviláglik, hogy a leggyakoribb kategória az elfogadás / egyetértés, azaz a beszélgetőpartnerek leginkább helyeslésüket fejezték ki a másik mondandója iránt. Ennek interakciós párja, a visszautasítás / egyet nem értés ugyanakkor ennél ritkábban fordul elő a korpuszban, a hatodik helyen található. Érdekes ugyanakkor megfigyelni, hogy a hárítás csak a tizenkettedik helyen szerepel, azaz a beszélgetőpartnerek inkább nyíltan felvállalják egyet nem értésüket, mintsem hogy kikerüljék a véleménynyilvánítást. Ezt valószínűleg magyarázhatja az is, hogy zárt közösségben készültek a hangfelvételek, a partnerek jól ismerik egymást, kicsi a köztük levő szociális távolság, ami együtt jár az udvariassági stratégiák alkalmazásának csökkenésével (Wolfson, 1988).

A 3-4., valamint a 7-8. leggyakoribb kategóriák is pozitív udvariassági stratégiát képviselnek, azaz a beszélgetőpartnerek közti szolidaritást hivatottak megerősíteni. Ugyanakkor a második leggyakrabban előforduló kategória a beszélő negatív véleményét fejezi ki, akár a beszédpartner, akár külső személy vagy tény iránt, a közvetlen figyelmeztetés vagy fenyegetés azonban igen ritkán található meg a korpuszban. Ez arra utal, hogy valószínűleg a partnerek inkább egy harmadik személy vagy külső tényezők iránti nemtetszésüket fejezik ki gyakrabban.

Megemlítjük azt is, hogy az ironia és antiironia eszközeivel viszonylag ritkán élnek élőbeszédben a beszélők, legalábbis a korpusz adatai szerint. Ugyanakkor az

Sorszám	Kategória	Gyakoriság
1.	elfogadás / egyetértés	10 659
2.	panasz / vád / kritika / sértés	3845
3.	kérés / parancs / kívánság	1875
4.	ígéret / ajánlat	1832
5.	figyelem felhívása	1566
6.	visszautasítás / egyet nem értés	1493
7.	dicséret / bók	1442
8.	üdvözlés / elköszönés	1161
9.	köszönetnyilvánítás	798
10.	bocsánatkérés	766
11.	irónia	493
12.	hárítás	299
13.	figyelmeztetés / fenyegetés	210
14.	antiirónia	24
Összesen		26 463

1. táblázat. Annotált pragmatikai egységek gyakorisága.

interaktív elemek (üdvözlés / elköszönés, valamint a figyelem felhívása) gyakori előbeszédi sajátságának mondhatók.

A pragmatikai szereppel bíró kifejezések leggyakoribb szavairól statisztikát is készítettünk, melyet az 1. ábra mutat be szófelhő formájában. Ebből kiviláglik, hogy elsődlegesen a köszönés, egyetértés, hezitáció és figyelemfelhívás szavai fordulnak elő. Ez részben összhangban áll a leggyakoribb annotált kategóriákkal, ugyanakkor arra is rávilágít, hogy feltehetőleg e beszédaktusoknak a leginkább korlátozott a szókincse, hiszen míg megkérni valakit vagy panaszkodni valamire sokféle nyelvi kifejezéssel lehetséges, addig például az üdvözlés vagy bocsánatkérés beszédaktusára csak viszonylag limitált számú szó és kifejezés létezik nyelvenként. Feltűnő még a diskurzusjelölők nagy száma is az ábrán, többek között a *hát*, *oké*, *egyébként* kifejezések is sűrűn szerepelnek a beszélt nyelvben.

4.2. Bizonytalansági annotáció

Kategória	Gyakoriság
weasel	7303
hedge	7166
feltételes	4885
doxasztikus	4131
peacock	2625
episztémikus	2209
kutatási	21
Összesen	28 340

2. táblázat. Bizonytalansági kategóriák gyakorisága.



1. ábra: A leggyakrabban használt szavak a pragmatikai kifejezésekben. (A „hz” a hezitálás, a „bs” pedig a beszédszándék jelölésére szolgál az annotációban.)

A 2. táblázat szemlélteti a bizonytalansági kategóriák gyakoriságát. Minden-
képpen említésre méltó, hogy a rangsor első két helyét a diskurzussztű bizony-
talanság két eleme, nevezetesen a weasel és hedge kategóriák foglalják el, azaz
inkább az élő diskurzusra jellemzők ezek a kategóriák. A feltételes kifejezések is
gyakoriak a korpuszban, emellett a hiedelmekre utaló kifejezések is szép számmal
fordulnak elő. Valószínűleg ez annak köszönhető, hogy a beszélgetések során a
partnerek sokszor fejezik ki, hogy saját véleményükről, elképzelésükről van szó,
nem általánosságban beszélnek. A kutatási kategória – egyáltalán nem meglepő
módon – szinte alig fordul elő a korpuszban.

A 2. ábrán láthatjuk a leggyakrabban használt bizonytalansági kifejezéseket. A doxasztikus kategóriában legkiemelkedőbb talán a *szerintem* és *gondolom* szavak szerepe, amelyek a beszélő véleményét fejezik ki. A *tudom* szó egyrészt előfordulhat episztemikus kifejezésekben *nem tudom*, másrészt doxasztikus kifejezésekben is *úgy tudom*, nem meglepő módon gyakran fordul elő a korpuszban. Gyakori weasel szónak számít az *izé*, ami valószínűleg élőbeszédi sajátosság, ilyen még a *nemtom* és a *tők* szó, melyek viszonylag ritkák más, bizonytalanságra annotált magyar korpuszokban.

A *kéne, lehetne, érted, tudod, hogyha, kicsit, annyira* stb. szavak mind a pragmatikai, mind a bizonytalansági szófélhőben előfordulnak, ami arra utalhat, hogy a pragmatikai és udvariassági kifejezésekben sűrűn használunk bizonytalansági kifejezéseket, ezzel enyhítve bizonyos beszédaktusok homlokzatfenyegető hatását a partnerre nézve.



4.3. A bizonytalanság és pragmatikai annotáció kapcsolata

Az adatokból kiviláglik néhány érdekesség. Először, a feltételes bizonytalanság leginkább a kérés / parancs / kívánság, illetve az ígéret / ajánlat beszédaktusokban fordul elő a leggyakrabban, valószínűleg a gyakran használt feltételes igealakoknak köszönhetően. Másodszor, a doxasztikus (a beszélő hiedelmeire utaló) bizonytalanság leggyakrabban a bocsánatkérés, visszautasítás és a hárítás beszédaktusokban tűnik fel, amit a *hisz* ige gyakori jelenléte magyarázhat: az *azt hittem/hiszem, hogy...* frázis sokszor enyhíti a homlokzatfenyegető aktus élet. Harmadszor, megjegyezzük, hogy a peacock kifejezések leginkább az ironikus, illetve antiironikus kifejezésekben találhatók a legnagyobb arányban: az (anti)ironikus megnyilatkozásokban sűrűn szerepelnek túlzások, így peacock kifejezések is a használt kifejezések és a mögöttes tartalom közti ellentét hangsúlyozására. Negyedszer, a hedge és weasel kategóriák esetében nem láthatunk különösebben eltérő használati gyakoriságot a pragmatikai kategóriák függvényében, viszont e kettő a leggyakoribb bizonytalansági kategória szinte mindegyik pragmatikai kategóriára nézve.

Kategória	hedge	weasel	peacock	episzt.	doxaszt.	felt.	kut.	össz
elfogadás/egyetértés	107	62	97	58	93	67	0	484
antiirónia	0	1	3	0	0	0	0	4
bocsánatkérés	37	34	16	12	73	24	0	196
figyelem felhívása	1	6	0	2	11	2	0	22
hárítás	24	9	0	9	42	6	0	90
panasz/vád/kritika/sértés	1199	987	689	184	526	539	1	4125
dicséret/bók	382	240	244	46	121	145	0	1178
üdvözlés/elköszönés	7	1	3	1	9	5	0	26
irónia	81	44	96	11	13	29	0	274
ígéret/ajánlat	233	121	25	61	37	171	0	648
visszautasítás/egyet nem értés	76	64	31	41	77	39	0	328
kérés/parancs/kívánság	114	101	26	21	48	219	0	529
köszönetnyilvánítás	12	0	2	0	0	0	0	14
figyelmeztetés/fenyegetés	32	21	18	11	3	25	0	110
Összesen	2305	1691	1250	457	1053	1271	1	8028

3. táblázat. A bizonytalanság előfordulása pragmatikai kifejezésekben.

5. Összegzés

A dolgozatban bizonyos pragmatikai és szemantikai sajátságokat vizsgáltunk magyar nyelvű, nagy méretű spontánbeszéd-korpusz (StaffTalk) alapján. A vizsgálati korpusz, amely egyedülálló a magyar nyelvű spontánbeszéd-adatbázisok körében, számos szemantikai és pragmatikai sajáttság kézi annotációióját tartalmazza. A korpuszt ezek alapján az annotációk alapján, elsősorban kvantitatív szempontból elemeztük a jelen dolgozatban. Bemutattuk a leggyakrabban használt pragmatikai jelenségeket, valamint a nyelvi bizonytalanság néhány élőbeszédi érdekességére is felhívtuk a figyelmet.

A jelen dolgozatban nem volt mód arra, hogy az egyes kategóriák elemeit kvalitatív és kvantitatív szempontból, alaposabban, a szakirodalmi megállapításokkal behatóan összevetve elemezzük. Tekintettel arra, hogy a korpusz annotációjának részletessége és a korpusz méretei nemzetközi szinten is kiemelkedőek, a kutatás következő lépéseként ezeket a vizsgálatokat tervezzük elvégezni. Ahogyan arra az eredmények tárgyalásában is igyekeztünk utalni, mindezek az elemzések számos fontos adalékot adhatnak a szemantikai és pragmatikai kutatásokhoz a jövőben.

Köszönetnyilvánítás

A korpusz létrehozását az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programja támogatta az ERC_CoG_2014_648693 sz. szerződésben, a kutatás vezetője Takács Károly.

Szabó Martina Katalin kutatásait részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal – NKFIH OTKA posztdoktori kiválósági programja (NKFI-azonosító: 132312) támogatta.

Szeretnénk megköszönni a korpusz annotátorainak kitartó és áldozatos munkáját.

Hivatkozások

- Andersson, L.M., Pearson, C.M.: Tit for tat? the spiraling effect of incivility in the workplace. *Academy of management review* 24(3), 452–471 (1999)
- Austin, J.L.: How to do things with words, vol. 88. Oxford university press (1975)
- Brown, P., Levinson, S.C.: Universals in language usage: Politeness phenomena. In: *Questions and politeness: Strategies in social interaction*, pp. 56–311. Cambridge University Press (1978)
- Calhoun, S., Carletta, J., Brenier, J.M., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D.: The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation* 44(4), 387–419 (2010)
- Cheng, W., Greaves, C., Warren, M.: The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). *ICAME journal* 29, 47–68 (2005)
- Crowdy, S.: Spoken corpus design. *Literary and Linguistic Computing* 8(4), 259–265 (1993)
- Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., Potts, C.: A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013)
- De Felice, R., Darby, J., Fisher, A., Peplow, D.: A classification scheme for annotating speech acts in a business email corpus. *Icane Journal* 37, 71–105 (2013)
- Deme, A., Markó, A.: Lengthenings and filled pauses in Hungarian adults’ and children’s speech. KTH Royal Institute of Technology (2013)
- Galántai, J., Pápay, B., Kubik, B.G., Szabó, M.K., Takács, K.: A pletyka a társas rend szolgálatában-az informális kommunikáció struktúrájának mélyebb megértéséért a computational social science eszközeivel. *Magyar Tudomány* 179(7), 964–976 (2018)
- Gulyás, A., Galántai, J., Szabó, M.K., Szebeni, Z.: A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai. In: *MSZNY 2018 - XIV. Magyar Számítógépes Nyelvészeti Konferencia*. pp. 317–330 (2018)
- Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The ATIS spoken language systems pilot corpus. In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990* (1990)
- Holmes, J., Stubbe, M.: Power and politeness in the workplace: A sociolinguistic analysis of talk at work. Routledge (2015)
- Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Van Ess-Dykema, C.: Automatic detection of discourse structure for speech recognition and understanding. In: *1997 IEEE*

- Workshop on Automatic Speech Recognition and Understanding Proceedings. pp. 88–95. IEEE (1997)
- Kane, J., Pápay, K., Hunyadi, L., Gobl, C.: On the Use of Creak in Hungarian Spontaneous Speech. In: ICPHS. pp. 1014–1017 (2011)
- Leech, G., McEnery, T., Weisser, M.: Spaac speech-act annotation scheme. University of Lancaster (2003)
- Leech, G.N.: Principles of pragmatics. Routledge (2016)
- Lennes, M., és mtsai: Segmental features in spontaneous and read-aloud Finnish. Phonetics of Russian and Finnish general description of phonetic systems: experimental studies on spontaneous and read-aloud speech (2009)
- Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous Speech Corpus of Japanese. In: LREC. pp. 947–9520. Citeseer (2000)
- Müller, C., Strube, M.: Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (szerk.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 197–214. Peter Lang, Frankfurt a.M., Germany (2006)
- Neuberger, T., Gyarmathy, D., Grácsi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: International Conference on Text, Speech, and Dialogue. pp. 424–431. Springer (2014)
- Oostdijk, N.: The Spoken Dutch Corpus. Overview and First Evaluation. In: LREC. pp. 887–894. Athens, Greece (2000)
- Pápay, K., Szeghalmy, S., Szekrényes, I.: Hucomtech multimodal corpus annotation. Argumentum 7, 330–347 (2011)
- Paul, G.H., és mtsai: Logic and conversation. Syntax and semantics 3, 41–58 (1975)
- Reichel, U.D., Mády, K.: Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech (2013), <http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-18043-4>
- Rogers, P.S., Lee-Wong, S.M.: Reconceptualizing politeness to accommodate dynamic tensions in subordinate-to-superior reporting. Journal of Business and Technical Communication 17(4), 379–412 (2003)
- Roque, A., Tsuetaki, A., Sarathy, V., Scheutz, M.: Developing a corpus of indirect speech act schemas. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 220–228 (2020)
- Searle, J.R.: Indirect speech acts. In: Speech acts, pp. 59–82. Brill (1975)
- Szabó, M.K., Vincze, V., Ring, O., Üveges, I., Vit, E., Samu, F., Gulyás, A., Galántai, J., Szvetelszky, Zs., Bodor-Eranus, E.H., Takács, K.: StaffTalk: magyar nyelvű spontán beszélgetések korpusza. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)
- Van Bael, C., Baayen, R.H., Strik, H.: Segment deletion in spontaneous speech: a corpus study using mixed effects models with crossed random effects. In: INTERSPEECH. pp. 2741–2744 (2007)
- Vincze, V.: Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14–18, 2013. pp.

- 383–391. Asian Federation of Natural Language Processing / ACL (2013), <https://www.aclweb.org/anthology/I13-1044/>
- Vincze, V.: Uncertainty detection in Hungarian texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1844–1853. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014), <https://www.aclweb.org/anthology/C14-1174>
- Vincze, V.: Detecting uncertainty cues in Hungarian social media texts. In: Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM). pp. 11–21. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), <https://www.aclweb.org/anthology/W16-5002>
- Wolfson, N.: The bulge: A theory of speech behaviour and social distance. In: Fine, J. (szerk.) Second Language Discourse: A Textbook of Current Research, pp. 21–38. Ablex, Norwood, N.J. (1988)
- Zhu, X., Penn, G.: Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. pp. 197–200 (2006)

Egy nyílt forráskódú magyar időpont-egyeztető chatbot

Nagy Soma Bálint, Herdinai Viktor, Farkas Richárd

Szegedi Tudományegyetem, Informatika Intézet
6720, Szeged, Árpád tér 2.
{nagysomabalint,herdinai.viktor98}@gmail.com,
rfarkas@inf.u-szeged.hu

Kivonat A dialógusrendszerek napjainkra a nyelvtechnológia egyik fontos területévé váltak. Jelen munkánkban egy általunk fejlesztett egyszerű magyar nyelvű szöveges chatbotot mutatunk be. A közzétett chatbot egy konzultációs alkalomra történő időpontfoglalást tesz lehetővé. Publikálunk továbbá egy dátum és időpont entitások strukturált kinyerésére fejlesztett könyvtárat, ami az időpontegyeztetésre szolgáló chatbotunk egy kulcsfontosságú része. Munkánk során feltérképeztük a magyar nyelvű chatbot fejlesztésre elérhető eszközöket. Jelen munka az első lépése annak, hogy megértsük, milyen nyelvtechnológiai kihívások állnak az üzleti chatbotok előtt, megismerjük a jelenkori dialógusrendszerek határait.

1 Bevezetés

A chatbotok elképesztő popularitásra tettek szert az elmúlt évtizedben. Míg korábban elsősorban kereskedelmi célú chatbot fejlesztés folyt, az elmúlt néhány évben a számítógépes nyelvészetet is elérte a chatbot láz. Míg korábban csak egyszerű, szabályalapú chatbotokat fejlesztettek, az elmúlt években több ezer nyelvtechnológiai publikáció született, amik a dialógusrendszerek problémáit és az azokra adott megoldásokat tárgyalják.

A magyarul “beszélő” chatbotok piaca nemzetközi szinten mozog¹. Ezeknek az üzleti chatbotoknak a működése azonban nem publikus, és a magyar nyelvű dialógus rendszereket folytató chatbotok tudományos szakirodalma pedig nagyon ritkás.

Úgy gondoljuk, hogy amennyiben a magyar üzleti chatbotok bonyolultabb, természetes emberi interakciókat akarnak kezelni, akkor szükségük van magyar nyelvi jelenségek kezelésére alkalmas algoritmusokra, amelyek ma még nem minden esetben állnak rendelkezésünkre. Jelen munkánkat az első lépésnek szánjuk azon az úton, hogy megértsük, milyen számítógépes nyelvészeti problémák lépnek fel üzleti dialógus rendszerekben, majd ezekre megoldásokat adjunk.

Annak érdekében, hogy a jelenlegi chatbotok határait és továbbfejlesztési igényét megértsük, egy egyszerű chatbot alkalmazást fejlesztettünk, melynek célja, hogy időpontot egyeztessen valamilyen hivatalos(abb) környezetben. Alkalmazási esetként

¹ A teljesség és sorrendiség igénye nélkül néhány magyar chatbot: talkabot.net, roborobo.hu, hanga.clementine.hu

beszédalapú (telefonos) ügyfélszolgálatok mögötti dialógusrendszert vagy szöveges chatbotot (például szakdolgozó konzultációs időpontot egyeztet a témavezetőjével) tartjuk a szemünk előtt. Mivel telefonos beszédalapú interfész mögött is helyesen kell működnie a rendszernek, ezért ebben a munkánkban csak folyószöveges kommunikációra koncentrálnunk (és például nem vizsgáljuk a chatbotok által gyakran használt kattintással listaelem választás lehetőségét).

Egy másik célunk volt, hogy a rendelkezésre álló nyílt forráskódú megoldásokat feltérképezzük. Úgy gondoljuk ugyanis, hogy a magyar számítógépes nyelvészeti közösségnek nem szabad egyik dialóguságens-fejlesztésére kialakított piaci keretrendszer mellett sem elköteleznie magát, és hiszünk abban, hogy a jövőben létrehozandó magyar nyelvtechnológiai megoldásainkat is nyílt forráskóddal kell elérhetővé tenni.

Cikkünkben bemutatjuk a nyílt forrású technológiákat használó időpont-egyeztető chatbotunkat, valamint az ennek megvalósításához kulcsfontosságú szöveges dátum felismerő, illetve időpont szöveg generálására fejlesztett szabályalapú rendszerünket, amelyek a <https://github.com/szegedai/hun-appointment-chatbot> és <https://github.com/szegedai/hun-date-parser> címeken elérhetőek.

2 Kapcsolódó munkák

Legjobb tudomásunk szerint egyetlen tudományos publikáció született eddig, amely magyar nyelvű chatbotot mutat be: Kemény és Recski (2018) egy utazástervező Facebook chatbotot fejlesztett. Célul azt tűzték ki, hogy egy nagyon rövid kérdésre a megfelelő menetrendi választ kérdezzék le a BKK API használatával. Munkájuk számítógépes nyelvészeti vonatkozásai a nagyon tömör, nyelvileg helytelen kérdések megértésére irányultak. Ezzel szemben a mi időpont-egyeztető chatbotunk feltételezi a nyelvileg helyes bemenetet, és a több körben megvalósuló egyeztetésre fókuszál.

A Lara rendszer (Nagyfi, 2018) nyílt forráskódú, magyar chatbot fejlesztésére kidolgozott eszközök gyűjteménye, amely kiterjedt funkcionalitással rendelkezik. A Lara könyvtár képes felhasználói szándék meghatározására, bizonyos entitások kinyerésére, illetve számos olyan magyar nyelvre fejlesztett nyelvi megoldást tartalmaz, amely segítséget nyújt chatbotok fejlesztéséhez.

A chatbotunk elkészítésére végül nem a Lara-t, hanem az univerzálisabban használható Rasa könyvtárat alkalmaztuk, amely a Lara-val szemben elsősorban gépi tanulás alapú, és egy széleskörűen használható keretrendszert biztosít a fejlesztő számára. A dátumok és időpontok kinyerésére vizsgáltuk a Lara használatának lehetőségét, ami tapasztalatunk alapján pontosan ismeri fel ezeket az entitásokat, azonban mivel nem képes időintervallumok, valamint relatív módon megadott dátumok kinyerésére, végül erre a célra saját eszközt fejlesztettünk.

² Például dialogflow.cloud.google.com vagy aws.amazon.com/chatbot

3 Architektúra

A munkánk során egy automatizált, szövegalapú asszisztensek és chatbotok fejlesztésére készült keretrendszert, a Rasa-t (Bocklisch és mtsai., 2017) alkalmaztuk. A Rasa számos természetes nyelv feldolgozási feladat elvégzésére képes, Spacy-hez készült magyar nyelvű NLP modelleknek köszönhetően magyar nyelvvel is kompatibilis. Az időpont-egyeztetéshez dátum, illetve idő típusú entitások kinyerésére volt szükség. A Rasa által is támogatott névelem kinyerésén túl arra volt szükségünk, hogy ezeket az entitásokat strukturált formára hozzuk, az elérhető időpontokat tartalmazó adatbázissal össze tudjuk hasonlítani őket. Ennek a részfeladatnak a megoldására megvizsgáltuk az elérhető szoftvereket, végül azonban saját implementáció készítése mellett döntöttünk. Az általunk megvalósított könyvtár képes időpont és időintervallum entitások felismerésére és normalizálására, valamint a strukturált formában tárolt dátumokból és időpontokból hétköznapi beszédben használatos szöveges reprezentáció generálására, melynek segítségével a chatbottal folytatott párbeszéd hasonlóbbá válik a való életben folytatott beszélgetésekhez.

Az időpontegyeztető chatbotunk és a magyar nyelvű szöveges dátumokat kezelő kódjaink Apache 2.0, illetve MIT licensszel elérhetőek a <https://github.com/szegedai/hun-appointment-chatbot> és <https://github.com/szegedai/hun-date-parser> felületeken.

3.1 A Rasa keretrendszer

A Rasa egy nyílt forráskódú, gépi tanuló keretrendszer, amelynek célja dialógusalapú rendszerek fejlesztése. A Rasa képes gépi tanulás és előre definiált szabályrendszer alkalmazásával

- a felhasználó üzeneteiből a felhasználó szándékát meghatározni,
- kinyerni a szövegből a releváns entitásokat, valamint
- a beszélgetés irányát terelni.

A Rasa számos olyan NLP modellet tartalmaz, amelyre dialógusalapú rendszer fejlesztésekor szükség lehet. A keretrendszer tartalmazza a Rasa kutatói által publikált, általunk is alkalmazott mélytanuláson alapuló DIET (T. Bunk és mtsai., 2020) architektúra implementációját, ami felhasználói szándék és szövegben előforduló entitások egyidejű kinyerésére szolgál. A DIET egy multi-task tanuló, transzformer alapú architektúra. Fejlesztésekor fontos szempont volt a modularitás, ennek köszönhetően számos különböző, előre betanított szóbeágyazással képes együtt működni, illetve egyes részei (intent, illetve entitáskinyerés) külön is használhatók. A DIET a tanítómondatokon a fenti két feladat mellett egy nyelvi modellet is tanul, és ezen három komponens együttes hibájának a minimalizálására törekszik a tanulás során. Munkánk során az általuk bemutatott módon alkalmaztuk a DIET-et, azzal a kivétellel, hogy az entitáskinyerő lépést a DIET helyett mi valósítottuk meg.

Amellett, hogy a Rasa számos problémára kész megoldást kínál, lehetővé teszi bármilyen, fejlesztő által megvalósított feldolgozási lépés integrálását, a meglévő lépések széleskörű konfigurációját.

Azért választottuk a Rasa-t a chatbotunk alapjául, mert nyílt forráskódú, könnyen testreszabható és bő funkcionálisitással rendelkezik. A Rasa mellett kipróbáltuk a

Deeppavlov keretrendszer is, azonban ezt jelentősen nehezebben használhatónak találtuk, valamint döntő kritérium volt, hogy a Rasa könnyebben kompatibilissé tehető a magyar nyelvvel, a Deeppavlov esetén erre a rendszer részét képző multilingual BERT (Devlin és mtsai. 2018) használata adhatott volna részben megoldást.

3.2 Természetes nyelv feldolgozási lépések magyar nyelven a Rasa-val

Számos hasonló célra készült szoftverrel ellentétben a Rasa nyelvagnosztikus, az általa használt modellek nyelvtől függetlenül alkalmazhatók. Ahhoz, hogy nyelvspecifikus ismeretekkel rendelkezzen a létrehozott asszisztens, a Spacy természetes nyelv feldolgozást lehetővé tevő könyvtárat, és a hozzá készült nyelvspecifikus NLP modelleket kell használnunk. Mivel a Spacy jelenleg nem rendelkezik hivatalos magyar nyelvű modellekkel, csupán a tokenizálást támogatja, ezért magyar korpuszon tanított, Spacy-vel kompatibilis, függetlenül közzétett modelleket (Orosz Gy., 2019) használtunk. A könyvtár telepítése után a Rasa automatikusan eléri a magyar nyelvű korpuszon tanított CBOW vektorokat, névelem felismerő, illetve lemmatizáló modelleket. Jelenlegi architektúránknak ezek közül a tokenizálás, illetve a folytonos vektortérbeli szóbeágyazás a részei.

Az üzenetből a felhasználó szándékát a transzformeralapú, a Rasa munkatársai által publikált és implementált DIET osztályozót alkalmaztuk, amely egy intentek és entitások egyidejű meghatározására képes architektúra. Mivel a rendszerünk által felismert entitások csupán a dátumok, illetve időpontok voltak, melyre saját entitáskinyerő lépést implementáltunk, ezért a DIET-et csak a felhasználói szándék kinyerésére használtuk. A DIET architektúra képes sűrű és ritka szóbeágyazások együttes használatával javítani a felismert intent pontosságán. A magyar nyelvű Spacy CBOW vektorok szolgálták a sűrű vektorokat, a száznál kevesebb tanítómondaton képzett szózsák modell kimenete pedig a ritka vektorokat. Ezen tanítómondatokat az intentek detektáláshoz határoztuk meg. A BOW elkészültekor használt szavak száma kevés, azonban felhasználását mégis előnyösnek ítéltük, hiszen a felhasználóktól is várható időpontfoglaláskor gyakran használatos szavakat, kifejezéseket tartalmazza a tanítóadatbázisunk. Arra számítottunk, hogy a szótáron kívül eső szavak ritkák lesznek a használat során, így a BOW jó kiegészítés lehet a folytonos reprezentáció használata mellé

3.3 Felhasználói felület

Jelenleg a chatbottal CLI-n keresztül lehet beszélgetést folytatni, azonban a Rasa lehetőséget biztosít, hogy az elkészült chatbotot webhook-ként közzétegyük, így népszerű közösségi oldalak üzenetküldő szolgáltatásaival, valamint azonnali üzenetküldésre szolgáló applikációkkal integrálhassuk.

3.4 Szöveges dátumreprezentáció strukturált formára hozása

Ahhoz, hogy a felhasználó által igényelt időpontot összehasonlíthassuk az elérhető időpontok listájával, a természetes szövegben számos formában előforduló dátumra és időpontra utaló kifejezéseket strukturált formára kellett hoznunk. Ehhez két magyar nyelven elérhető, szabályalapú rendszert próbáltunk ki. Az első a Duckling (Facebook, 2016) volt, ami egy nyílt forráskódú, entitások elemzésére szolgáló szoftver. Tapasztalatunk alapján a Duckling nagyon jó pontossággal működik angol nyelven, azonban a magyar dokumentumokon nem képes dátumok robusztus felismerésére. Ennek oka az lehet, hogy angol nyelvre sokkal bővebb szabályrendszer áll rendelkezésre, valamint az, hogy a szoftver nem képes megfelelően kezelni a magyar nyelv agglutinációból adódó sajátosságait.

Ezt követően a Lara (Nagyfi R., 2018) nevű, speciálisan magyar nyelvre fejlesztett könyvtárat vizsgáltuk, amely tapasztalatunk szerint mind dátumok, mind időpontok kinyerésében jobban teljesített, mint a Duckling, azonban a Ducklinggal ellentétben nem képes időintervallumok, napszakok, napok nevével megnevezett dátumok értelmezésére. Vizsgáltuk a Lara által kezelt esetek lehetséges kibővítését, azonban felismertük, hogy a céljainkra egy olyan könyvtár felelne meg igazán, ami képes a dátumot és időt leíró szavakat időintervallumként értelmezni. Ezen tapasztalataink alapján úgy döntöttünk, hogy saját implementációt készítünk dátumok és időpontok strukturált formában való kinyerésére. Az elkészült szoftver Python nyelven készült, tervezésekor figyelmet fordítottunk arra, hogy amellet, hogy a saját céljainkra alkalmas legyen, egy mások által is könnyen felhasználható könyvtárat fejlesszünk.

3.5 Dátumok és időpontok természetes szöveges formára hozása

Ahhoz, hogy a chatbot válaszai a hétköznapi beszédben megszokotthoz hasonlóak lehessenek, a válaszokban megjelenő dátumokat és időpontokat strukturált reprezentációról természetes szöveggé alakítottuk. Mivel nem ismert számunkra erre létező megoldás, mi készítettünk implementációt a probléma megoldására. A megoldást jelentő programot a dátum és időpont kinyerő könyvtárunk részeként publikáljuk.

3.6 Az elérhető időpontok adatbázisa

Jelenleg a chatbot számára elérhető szabad időpontok listáját egy statikus fájlban tároljuk, a felhasználóval való időpont-egyeztetés során innen olvassuk ki azon időpontokat, amelyek a bot számára elfogadhatók, és amelyek közül a felhasználó kérésére ajánlatot tesz a bot. A jövőben természetesen ezt ki lehet majd cserélni valamilyen naptáralkalmazás API-n keresztüli hívására.

4 Eredmények

4.1 Magyar nyelvű természetes szövegben található dátumok kezelése

A korábban kifejtett okok miatt saját szoftverrel végezzük a dátumra, időintervallumra és időpontra utaló entitások kinyerését, valamint strukturált formára hozását. Munkánk során egy reguláris szabályokat alkalmazó, lehetőség szerint az írott természetes nyelvben előforduló jelenségeket kezelő dátumkinyerőt fejlesztettünk. A dátumkinyerő jelen verziója nem támaszkodik természetes szöveg feldolgozását segítő könyvtárakra, a néhány szükséges, szöveges előfeldolgozási lépést mi valósítottuk meg a program részeként.

Ahhoz, hogy széleskörűen használható legyen az elkészült szabályrendszer, valamint biztosítani tudja, hogy a felhasználó bármilyen formában hivatkozhasson a számára megfelelő időpontra, többfélelépp előforduló dátum- és időpontalak kezelését kellett megvalósítanunk:

- Dátumok és időpontok konkrét megjelenései a szövegben (például: *2020 november 10., dec. 2, három óra, 15:00, háromnegyed kettő után öt perccel* stb.)
- Névvel megnevezett napszakok (például: *reggel, délben, este* stb.)
- Az aktuális időponthoz képest relatív időpontok (például: *jövő kedden, holnap* stb.)

Ezeket az eseteket szabályokká képeztük le, figyelmet fordítva arra, hogy a szabályrendszerünk szükség esetén bővíthető legyen. A program működése során a felhasználó által kért bemenetre illesztjük az egyes szabályokat. A könyvtár szabályai igyekezzenek lefedni a természetes nyelvben dátum, illetve időpont leírására használt kifejezőmódokat. Amennyiben több szabály is illeszkedik a felhasználói inputra, úgy ezek megfelelő kombinációja képi a kimenetet. Az egyes esetekhez tartozó kimenetet, abban az esetben, amikor sikeresen nyertünk ki a szövegből időintervallumot, egy kezdő- és egy végdátum alkotja.

Mivel a strukturált dátumformátum nem engedi meg, hogy bármilyen információ hiányozzon, ezért ha a felhasználó nem pontosított minden részletet a bemenetként szolgáló szövegben, egyszerű feltevések használatával pótoljuk a hiányzó mezőket. (Például: a “kedden” bemenet esetén feltételezzük, hogy az adott heti keddre utal a felhasználó, hiányzó év esetén az aktuális évet feltételezzük, csak időpont megadása esetén az aktuális napot, stb.)

Strukturált formából szöveget szintén szabályok alkalmazásával készítünk. Mivel egy dátumra többféleléppen is lehet hivatkozni (az adott dátumtól számított távolság, konkrét időpont, stb.), ezért ahol lehetséges, több érvényes jelöltet generálunk, melyek mind az adott időpontot írják le. Mivel a természetes nyelvben általában nem másodpercre pontosan fejezzük ki az időpontokat, ezért mi is biztosítjuk annak lehetőségét, hogy a felhasználó meghatározhassa, hogy milyen pontosságot fejezzen ki a generált szöveg.

Annak érdekében, hogy ne tartalmazzon redundáns információt a generált szöveg, és így a természetes beszédben használthoz hasonló kimeneteket képezhessünk, csak azokat a dátum által hordozott információkat képeztük le explicit módon szöveggé, amelyek feltétlenül szükségesek a dátum beazonosításához. A dátumkinyerőnél a

hiányos dátumrészekre alkalmazott feltevéseinket itt is érvényesnek tekintettük. (Például egy dátumhoz generált szöveg nem szükségszerűen tartalmazza az évet, ha aktuális évi dátumról van szó.)

4.2 A chatbot által kezelt scenáriók

A chatbot képes több forgatókönyvet is kezelni, ezekhez elsősorban felhasználói szándékokat kellett definiálni. A jelenleg definiált intentek:

- üdvözlés,
- időpontfoglalási kérés,
- dátum megadása az igényelt időponthoz,
- igényelt idő megadása az időponthoz,
- elérhető időpont kérése a bottól, valamint
- a beszélgetés lezárását igénylő intent.

Ahhoz, hogy a felhasználói szándékokat felismerjük, tanító mondatokat kell megadni a Rasa-nak, arra, hogy milyen lehetséges felhasználói üzenetek tartoznak az egyes intentekhez. Tapasztalataink alapján ilyen alacsony számú intentnél 7-10 mondat esetén már megfelelő arányban azonosítja a helyes szándékot, de minél többet adunk meg, annál alacsonyabb a hibaarány. A chatbot tesztelése során azonosítottunk olyan lehetséges eseteket, amikor nem voltunk elégedettek a predikált szándékkal, ilyen esetben a tanítóhalmazt kibővítve a hibásan értékelt üzenetekhez hasonlókkal azt tapasztaltuk, hogy újbóli tanítás után a chatbot könnyen adaptálódott. Hasonló hatás érhető el a Rasa részét képező interaktív tanulást biztosító lehetőséggel, amelyet a chatbot jelenlegi verziójában nem alkalmaztunk. A chatbot a szándék felismerése után a beszélgetést ún. akciók végrehajtásával folytatja. Az akciók lehetnek konstans válaszok, de bonyolultabb esetben egyedi akciók implementálására is lehetőség van, ilyenkor egy meghatározott felépítésű osztály megvalósításával vihetünk komplexebb logikát a chatbot viselkedésébe. A chatbotunkhoz három egyedi akciót valósítottunk meg:

- Igényelt időpont összevetése a szabad időpontokkal, slot-ba mentése, visszajelzés arról, hogy az igényelt időpont megfelelő-e
- Ajánlatadás a szabad időpontok közül
- Slotok kiürítése

Azt, hogy mely intent után melyik akció hajtódjon végre, szabályok határozzák meg. Néhány egyszerű szabályt alkalmaztunk, további feltételek megadása nélkül, melyek leképzést adnak meg az összes intent és egy vagy több akció között.

A bot tanítása előtt lehetőség van „sztorik” megadására. Egy sztoriban egy lehetséges beszélgetést adunk meg absztrakt szinten, vagyis azt, hogy a beszélgetés során mely intentek kerülnének felismerésre, mely akciók futnának le, ezzel segítve a dialógusmodell tanulását. Fontos megjegyezni, hogy a sztorikban szereplő egymást követő intentek és akciók nem mondhatnak ellent a szabályokban leírtaknak, különben a tanítás sikertelen lesz. Több sztorit is megadtunk, amelyek forgatókönyvek egyes egyedi akciók használatára. Példa egy sztorira, ahol a dátum és az időpontot is megfelelő a chatbotnak:

```
- story: happy path
steps:
- intent: greet
- action: utter_greet
- intent: appointment_request
- action: validate_appointment
- intent: say_date
- slot_was_set:
  - date: "2021-02-17"
- action: validate_appointment
- intent: say_time
- slot_was_set:
  - time: "13:00"
- action: validate_appointment
- intent: request_end
- action: utter_bye
```

Látható, hogy egy sztori több valós beszélgetést lefed, hiszen a fenti sztori csak absztrakt szinten definiálja a beszélgetés folyását.

4.3 Időpont egyeztetés több körben

Olyan időpontegyeztetési folyamatokat modelleztünk, ahol egy tágabb időintervallumból kiindulva, a felhasználó és a bot a számukra alkalmas időtartamot néhány lépésben egyetlen konkrét időpontra szűkítjük, ezért ahol lehetséges, a fejlesztett rendszer időintervallumok kinyerésére törekszik a bemenetként szolgáló természetes szövegből.

Miután az opcionális kölcsönös üdvözlés megtörtént, a felhasználó jelezheti időpontfoglalási szándékát. Ha ez az üzenet nem tartalmazza a szükséges információkat a kívánt időpontra vonatkozóan, egy többfordulós folyamat kezdődik, amely során a felek kölcsönösen megfelelő időpontot keresnek.

Ha a felhasználó nem adta meg az időpontfoglalási szándékával együtt a kívánt dátumot, a chatbot rákérdez, hogy mikor felelne meg a felhasználónak. Ezután a beszélgetés szerteágazhat a válaszok alapján. Az említett dátumot a bot elfogadhatja, vagy elutasíthatja, dátum-ajánlatkérés esetén pedig két dátumot javasol. A felhasználó által kért dátum validálása a három akciónk egyikével történik, ezzel ellenőrizzük, hogy az adatbázisban szerepel-e az adott dátum, amennyiben igen, egy slotban eltároljuk, különben egy másik dátumot kérünk a felhasználótól. A dátumot tartalmazó slot beállítása után a bot “továblép” az időpont rögzítésére. A dátumhoz hasonlóan itt is az elfogadás, elutasítás, ajánlás esete következhet be. Az egyeztetés bármely része alatt a felhasználó kezdeményezheti, hogy a már megbeszélt dátum és időpont helyett találjanak egy másikat. Ennek hatására a slot-ban tárolt érték törlődik, megfelelő dátum esetén az új kerül beállításra. Sikeres időpontfoglalás után a chatbot megerősítésképp visszaolvassa a slotokból a rögzített időpontot mondatba foglalva.

4.4 Példa a chatbot működésére

Az alábbiakban a chatbot működését szemléltetjük a beszélgetés során háttérben zajló folyamatok magyarázatával kiegészítve. A beszélgetés tartalmaz példát a felhasználó

által ajánlott, a chatbot számára megfelelő, valamint nem megfelelő időpontokra, a chatbot által ajánlott időpontra, a felhasználó kérésére történő időpontfelülírássra. A beszélgetésben megfigyelhető, hogy a chatbot képes a beszélgetésből az időpont megállapításához szükséges dátum, illetve idő típusú entitások robusztus kinyerésére.

Felhasználó:	Szia	<i>(üdvözlés intent detektálva)</i>
Chatbot:	Szia! Miben segíthetek?	
Felhasználó:	Időpontot szeretnék foglalni	<i>(időpont foglalás intent detektálva, időpont nem nyerhető ki az üzenetből)</i>
Chatbot:	Okés. Mikor lenne jó?	
Felhasználó:	Te mikor érnél rá?	<i>(időpont ajánlást kérő intent detektálva)</i>
Chatbot:	Legközelebb ma és holnap érek rá.	<i>(elérhető nap ajánlása az időpontok listája alapján)</i>
Felhasználó:	Akkor találkozunk <u>kedden</u>	<i>(entitás detektálása, normalizálása, összevetése az időpont listával, slot feltöltése az entitás értékével, vizsgált intervallum: aktuális hét kedd, a beszélgetést követő nap)</i>
Chatbot:	Ráérek holnap. Mikor lenne jó aznap?	
Felhasználó:	Igazából egész <u>délelőtt</u>	<i>(entitás detektálása, normalizálása, összevetése az időpont listával, vizsgált intervallum: kedd 10-12 óra)</i>
Chatbot:	Sajnos nem érek rá ekkor... Egy másik időpont esetleg?	
Felhasználó:	Akkor amikor te ráérsz	<i>(időpont ajánlást kérő intent detektálva)</i>
Chatbot:	Holnap ráérek reggel 9-kor és délután 1-kor.	<i>(elérhető időpont ajánlása az időpontok listája alapján)</i>
Felhasználó:	Találkozunk <u>reggel kilenckor</u>	<i>(entitás detektálása, normalizálása, összevetése az időpont listával, slot feltöltése az entitás értékével, vizsgált intervallum: kedd 9 óra)</i>
Chatbot:	Rendben, akkor holnap reggel 9-kor találkozunk.	
Felhasználó:	Lehetne inkább a <u>szerda</u>?	<i>(időpont felülírását kérő intent detektálva, entitás kinyerés, normalizálás, összevetés az elérhető időpontokkal, slot feltöltése, vizsgált intervallum: aktuális szerda)</i>
Chatbot:	Ráérek ezen a héten szerdán. Mikor lenne jó aznap?	
Felhasználó:	Bármikor	<i>(időpont ajánlást kérő intent detektálva)</i>
Chatbot:	Ezen a héten szerdán ráérek délután 2-kor.	
Felhasználó:	Rendben, legyen <u>14 órakor</u>	<i>(entitás kinyerés, összehasonlítás az elérhető időpontokkal, slot feltöltés, vizsgált intervallum: szerda 14 óra)</i>
Chatbot:	Rendben, akkor ezen a héten szerdán délután 2-kor találkozunk.	

Felhasználó: **Köszí** *(elköszönés intent detektálva)*
Chatbot: **Viszlát!**

Egy lehetséges alternatív út az üdvözlés után:

Felhasználó: **Időpontot szeretnék foglalni vasárnapra** *(foglalás intent, időpont kinyerés, vizsgált intervallum: aktuális vasárnap)*
Chatbot: **Sajnos nem érek rá akkor... Egy másik esetleg?**
Felhasználó: **Mikor vagy szabad?** *(időpont ajánlást kérő intent detektálva)*
Chatbot: **Legközelebb ma és holnap érek rá.**

Miként korábban is utaltunk rá, vannak olyan esetek, amelyekben a chatbot nem kezeli megfelelően a felhasználói inputot, előfordulhat, hogy hibásan vagy egyáltalán nem detektál intentet. Erre részben megoldásként szolgált az, hogy a chatbot olyan esetekben, amikor minden lehetséges intenthez tartozó konfidencia érték alacsony, egy alapértelmezett választ ad a felhasználónak, melyben arra kéri, hogy üzenetét ismétlje meg más formában.

Példák alacsony konfidencia értékre:

Chatbot: **Okés. Mikor lenne jó?**
Felhasználó: **Most éppen ráérek.** *(alacsony konfidencia érték, fallback intent)*
Chatbot: **Elnézést, nem teljesen értettem. Át tudnád fogalmazni?**
Felhasználó: **Január 13-án ráérek.** *(foglalás intent, entitás kinyerés, vizsgált intervallum: aktuális nap)*
Chatbot: **Ráérek ma. Mikor lenne jó aznap?**

Felhasználó: **Szia** *(üdvözlés intent detektálva)*
Chatbot: **Szia! Miben segíthetek?**
Felhasználó: **Programozó szeretnék lenni.** *(nagyon alacsony konfidencia érték, fallback intent)*
Chatbot: **Elnézést, nem teljesen értettem. Át tudnád fogalmazni?**

5 Tapasztalatok és jövőbeli tervek

Jelenlegi tapasztalataink alapján a nyílt forrású Rasa keretrendszer alkalmas lehet komplexebb dialógusrendszerek megvalósítására is. Az egyedi akciók alkalmazásával könnyedén lehet a felhasználó kérésére az entitásoktól függő, dinamikus válaszokat adni. Használatukkal összeköthető a chatbot – API-kon keresztül – adatbázisokkal és egyéb webes szolgáltatásokkal.

A dátumfelismerés és a szöveges dátum generálással kapcsolatban azt tapasztaltuk, hogy a céljainkra már viszonylag egyszerű szabályrendszer is megfelelő, természetesen ahogyan a szabályok számát növeltük, úgy nőtt a rendszer robusztussága is.

A chatbotunk jelenleg nem képes bonyolultabb nyelvi szerkezetek feldolgozására. Eddigi tapasztalataink alapján, a legfontosabb szükséges magyar számítógépes nyelvészeti megoldások, amelyekre szükség lesz bonyolultabb szerkezetek megértéséhez:

- negáció detektálása (például: *“szerdán nem érek rá”*)
- módosító szerkezetek kezelése (például: *“szerda kivételével bármelyik nap jó”*)
- helytelen írásmódú, hiányos válaszok kezelése

A Rasa számos olyan képességgel rendelkezik, amelyeket a chatbotunk jelen verziójában nem használtunk, azonban a későbbiekben segítségünkre lehetnek, mint az interaktív tanulás a dialógusmodell fejlesztésére, tesztesetek megadása a dialógusrendszerre, vagy az integrációs lehetőségek közösségi platformokon.

A későbbiekben terveink közt szerepel a chatbot viselkedésének és teljesítményének kiértékelése emberi kérdőívek kitöltésével (Deriu és mtsai, 2020), valamint a chatbot meglévő működésének kibővítése, új funkciók hozzáadása. Ilyen funkció lehet, hogy a chatbot önmagától, külön kérés nélkül is képes legyen időpontot ajánlani. A jelenlegi statikus elérhető időpontok tárolását is leváltanánk egy valódi adatbázisra, ezzel a beszélgetés során egyeztetett időpontot foglaltként tudnánk megjelölni, elkerülve azt az esetet, hogy egy időpontot többen is lefoglaljanak. Továbbá a hitelesebb felhasználói élmény érdekében bővítenénk a válaszsablonok számát, így elkerülve az ismétlődő szerkezetű válaszokat a chatbot részéről.

Köszönetnyilvánítás

Farkas Richárd kutatási munkáját a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatta a 2018-1.2.1-NKP-2018-00008 azonosítójú projekt keretében.

Nagy Soma Bálint és Herdinai Viktor munkáját a "Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein" című, EFOP-3.6.3-VEKOP-16-2017-0002 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

A publikációban szereplő kutatást az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

Hivatkozások

- T. Bocklisch, J. Faulkner, N. Pawlowski, A. Nichol: Rasa: “Open Source Language Understanding and Dialogue Management”. *CoRR* abs/1712.05181 (2017)
- Burtsev M. és mtsai. “DeepPavlov: Open-Source Library for Dialogue Systems” Proceedings of ACL 2018, System Demonstrations. (2018)
- Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, Mark Cieliebak: “Survey on evaluation methods for dialogue systems” Artificial Intelligence Review (2020)
- J. Devlin és mtsai. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. *CoRR* abs/1810.04805 (2018)
- Orosz Gy., “Hungarian models for spaCy”. <https://github.com/oroszgy/spacy-hungarian-models>. (2019).
- Facebook, “Duckling”. <https://github.com/facebook/duckling>. (2016)
- Kemény Boldizsár, Recski Gábor: “Természetes nyelvi interfész menetrend- és utazástervező szolgáltatásokhoz” Magyar Számítógépes Nyelvészeti konferencia (2018)
- Nagyfi R, “Lara” <https://github.com/sedthh/lara-hungarian-nlp>. (2018)
- T. Bunk, D. Varshneya, V. Vlasov, and A. Nichol: “DIET: Lightweight Language Understanding for Dialogue Systems”. (2020)

POSZTER, LAPTOPOS BEMUTATÓ

StaffTalk: magyar nyelvű spontán beszélgetések korpusza

Szabó Martina Katalin^{1,2}, Vincze Veronika³, Ring Orsolya¹, Üveges István^{2,4},
Vit Eszter^{5,6}, Samu Flóra^{5,6}, Gulyás Attila¹, Galántai Júlia⁵, Szvetelszky
Zsuzsanna¹, Bodor-Eranus Eliza Hajnalka¹, Takács Károly^{5,1}

¹Társadalomtudományi Kutatóközpont, CSS-RECENS
1097 Budapest, Tóth Kálmán utca 4.

²Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.

³MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.

⁴Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
6722 Szeged, Egyetem utca 2.

⁵Linköpingi Egyetem, The Institute for Analytical Sociology
601 74 Norrköping, Svédország

⁶Budapesti Corvinus Egyetem
1093 Budapest, Fővám tér 8.

{martina,vincze}@inf.u-szeged.hu

{Szabo.Martina,Ring.Orsolya,Gulyas.Attila,Bodor-Eranus.Eliza}@tk.hu

{uvegesistvan898,szvetelszky}@gmail.com

{eszter.vit,flora.samu,julia.galantai,karoly.takacs}@liu.se

Kivonat A cikkben bemutatjuk a StaffTalk nevű, nagy méretű, kézzel annotált korpuszt, mely magyar nyelvű spontán beszélgetéseket tartalmaz. A korpusz létrehozásával elsősorban ahhoz szerettünk volna vizsgálati anyagot teremteni, hogy zárt közösségeken belül az informális kommunikáció és a megbecsültség hogyan befolyásolja a közösség működését és normarendszerét. A munka első lépéseként a hanganyagokat legépeltettük, amelynek során a verbális információn túl egyéb, nem verbális információk megjelölésére is megkértük az annotátorokat. A legépelt hanganyagokat ezt követően három szinten annotáltuk: a beszélgetésekben megjelenő pletykát, beszédaktusokat és egyéb pragmatikai jegyeket, valamint bizonytalanságra utaló szavakat egyaránt megjelöltünk. Mindezeknek a sajátságoknak köszönhetően a kiinduló kutatási kérdéssel összefüggésben, valamint azon túl is a korpusz sokféle pragmatikai szempontú elemzés elvégzésére is alkalmassá vált.

Kulcsszavak: korpusz, spontán beszéd, kézi annotálás, pragmatika, szemantika, pletyka

1. Bevezetés

Manapság a társadalomtudományok területén is egyre népszerűbbé válnak a korpuszalapú, illetve számítógépes nyelvészeti eszközöket alkalmazó vizsgálatok. Je-

len kutatási programunk arra a kérdésre keresi a választ, hogy zárt közösségekben belül az informális kommunikáció és a megbecsültség hogyan befolyásolja a közösség működését és normarendszerét, és a kutatás során a pletyka mint diskurzus vizsgálatára fókuszál. A pletyka ugyanis jelentős szerepet játszik az interperszonális informális kommunikációban, korábbi kutatások szerint ezeknek legalább a felét, de akár kétharmadát is lefedheti (Dunbar, 1996, 2004; Foster, 2004).

A kutatás keretében okosórák segítségével hangfelvételeket készítettünk zárt közösségekben. Az itt bemutatott kutatási fázisban egy iskola oktatói karának diskurzusait rögzítettük, majd e felvételek leiratozását és annotálását végeztük el.

A tanulmányban részletesen bemutatjuk a hangfelvételek keletkezési körülményeit, az anyagok feldolgozási módszereit és eszközeit, valamint a StaffTalk korpusz alapvető statisztikai adatait.

2. Kapcsolódó irodalom

Az alábbiakban röviden áttekintjük a spontán beszélt nyelvi adatbázisokra, valamint a korpuszban alkalmazott annotációs szintekre vonatkozó szakirodalmat.

2.1. Beszélt nyelvi korpuszok

A különböző, így többek között a társadalomtudományi és a nyelvészeti (például pragmatikai) tárgyú kutatások egyik legfontosabb vizsgálati eszközét a számítástechnikai eszközökkel elemezhető formátumú szövegtörzsek jelentik. A korpuszok között írott és beszélt nyelvi szövegtörzseket is találunk, azonban a legtöbb létező korpusz az írott nyelvet reprezentálja (McEnery, 2012). Ennek talán a legfontosabb oka az, hogy a beszélt nyelvi anyag feldolgozására jelenleg sokkal kevesebb eszköz áll a rendelkezésünkre, mint az írott nyelvi adatok kezelésére (Galántai és mtsai, 2018).

Az elmúlt évtizedekben több spontán beszélt nyelvi korpusz keletkezett több nyelven is (Crowdy, 1993; Hemphill és mtsai, 1990; Maekawa és mtsai, 2000; Oostdijk, 2000; Van Bael és mtsai, 2007), amelyek között találunk agglutináló nyelveket reprezentáló korpuszokat is (Neuberger és mtsai, 2014), például török (Mengusoglu és Deroo, 2001) és finn (Seppänen és mtsai, 2003).

Ugyanakkor a korpuszok növekvő száma ellenére még mindig csak néhány van, amely hangzó szövegeket tartalmaz, azok gépelt leírataival együtt. Ez az átírási eljárás magas munkaerő- és költségigényével magyarázható. Különösen csekély a magyar beszélt nyelvet reprezentáló korpuszok száma, és ezek is többségükben felolvasott szövegeket tartalmaznak (Gósy, 2013). Az első magyar nyelvű beszélt korpusz a 20. század elején keletkezett (Neuberger és mtsai, 2014). Az elmúlt évtizedekben készült beszédadatbázisok többsége rögzített olvasott beszédet, vezetett történetmondást vagy irányított beszélgetéseket tartalmaz. A magyar telefonbeszéd-adatbázis (MTBA) egy beszédkorpusz, amely 500 alany által telefonon rögzített olvasott szövegből áll. Úgy tervezték, hogy támogassa a

beszédtechnológia területén végzett kutatásokat és fejlesztéseket (Vicsi és mtsai, 2002). Az úgynevezett Kivi korpusz (Kugler, 2015) különféle, videón látott történetek elmeséléseit tartalmazza, míg a Budapesti Szociolingvisztikai Interjú 250 adatközlő interjújából áll (Várad, 2003). A HuComTech multimodális korpusz körülbelül 50 órányi video- és hangfelvételt tartalmaz 111 formális (szimulált állásinterjú) és 111 informális, de irányított párbeszédből (Pápay és mtsai, 2011). A szövegek létrejöttének körülményei, valamint a szövegek feldolgozási módja miatt azonban a fentebbi korpuszok nem támogatják a magyar spontán beszéd nyelvi sajátosságainak kutatását.

Legjobb tudomásunk szerint jelenleg három korpusz van, amely a magyar spontán beszédet kívánja reprezentálni, ez a Budapesti Egyetemi Kollégiumi Korpusz (BEKK) (Bodó és mtsai, 2017), a BEszélt nyelvi Adatbázis (BEA) (Gósy, 2013), valamint a HuTongue (Galántai és mtsai, 2018). Végezetül, meg kell még említenünk a CHILDES adatbázisból elérhető magyar gyereknyelvi korpuszokat is (Babarczy, 2009).

Ugyanakkor, a fenti korpuszok saját vizsgálataink elvégzésére nem voltak alkalmasak. Egyrészt, a BEKK esetében az interakciókat a résztvevők saját telefonjaikon rögzítették, ezért szelektív társalgásokat tartalmaz, ami nem reprezentálja tökéletesen az élőbeszédet. Másrészt, a BEA korpusz létrehozóinak fő célja az volt, hogy fonetikai, és nem szemantikai vagy pragmatikai vizsgálatokat tegyen lehetővé: így alakították ki a korpuszban alkalmazott annotációt. Az elmondottak okán a BEA-korpusz társadalomtudományi tárgyú kutatásokra csupán korlátozottan alkalmazható. Végül, az ún. HuTongue korpuszt (Galántai és mtsai, 2018) csupán félig (vagy részlegesen) spontánnak tekinthetjük, mivel egy szórakoztató jellegű tévéműsor céljaira készültek a felvételek, és, bár a társalgások a legtöbbször nem voltak kívülről kérdésekkel vagy témameghatározásokkal irányítva, a szövegek keletkezési körülményei (a résztvevők motivációi, valamint az időnkénti rendezői irányítás) befolyásolhatták a beszélői megnyilatkozásokat.

2.2. Pletykára annotált korpuszok

A pletykadiskurzusok elemzését, valamint társadalmi szerepét a korábbi kutatások többnyire kvalitatív vagy kérdőíves kvantitatív módszerekkel végezték (Eckhaus és Ben-Hador, 2019), írott szövegeken (Mitra és Gilbert, 2012) vagy megfigyelésen keresztül (Dunbar, 2004). A spontán előbeszéd belüli pletyka feltárására ezidáig kívül kevés korpusz és empirikus eredmény állt rendelkezésre (tudomásunk szerint kizárólag Robbins és Karan (2020)). Egyetlen magyar nyelvű korpuszról van tudomásunk (a fentebb említett HuTongue-ról), amely kifejezetten a pletyka természetének spontán beszéd belüli vizsgálatára irányul (Gulyás és mtsai, 2018; Szabó és Galántai, 2017), azonban a jelen tanulmányban bemutatott korpusz több szempontból jelentősen gazdagítja a pletyka vizsgálatának lehetőségeit a HuTongue korpuszhoz képest.

2.3. Pragmatikai annotációt tartalmazó korpuszok

Bár, amint azt korábban a 2.1. fejezetben emítettük, a nemzetközi irodalomban egyre több spontánbeszéd-adatbázissal találkozni, mind a nemzetközi, mind a hazai spontánbeszéd-vizsgálatokra jellemző, hogy azok alapvetően fonetikai, illetve akusztikus sajátságok elemzésére irányulnak (pl. (Kane és mtsai, 2011; Reichel és Mády, 2013; Deme és Markó, 2013; Lennes és mtsai, 2009; Zhu és Penn, 2006)). Ugyanakkor azt, hogy olyan szemantikai–pragmatikai sajátságokról, mint például bizonyos beszédaktusok, illetve a nyelvi udvariasság különböző formái, pontos és valós képet kaphassunk, nagy méretű, megfelelően annotált spontánbeszéd-korpuszokra van feltétlenül szükség.

A nemzetközi korpuszok közül a legtöbb, amely pragmatikai annotációt is tartalmaz, telefonbeszélgetésekből áll. Így például a brit Telecom 1200 telefonbeszélgetéséből készült OASIS korpusz beszédaktus-szintű annotációt tartalmaz (Leech és mtsai, 2003). A Switchboard korpusz, amelyet több különböző sajtóság mentén is annotáltak, szintén tartalmazza a beszédaktusok tageit is (Calhoun és mtsai, 2010). A dialógusaktusok jelentősen több típusát annotálták a fentebbi Switchboard korpusz egy részén (Jurafsky és mtsai, 1997).

Ami a magyar nyelvet illeti, jelenleg egyetlen olyan magyar, beszélt nyelvi korpuszról van tudomásunk (HuComTech), amely diskurzus szintű annotációt is tartalmaz, azonban ez az annotáció mindössze négy sajtóságra terjed ki (turn-taking, turn-giving, backchannel, turn-keeping) (Pápay és mtsai, 2011).

2.4. Bizonytalanságra annotált korpuszok

A bizonytalanságot jelző nyelvi elemek vizsgálata intenzív kutatási területnek számít a számítógépes nyelvészetben, meg kell jegyeznünk ugyanakkor, hogy az eddigi vizsgálatok néhány kivételtől eltekintve az angol nyelv köré csoportosulnak, és elsődlegesen újsághíreket, biológiai publikációkat vagy orvosi dokumentumokat, illetve Wikipedia-szócikkeit elemezik (vö. Szarvas és mtsai (2012); Kim és mtsai (2008); Sauri és Pustejovsky (2009)). Tudomásunk van mindemellett két magyar nyelvű, bizonytalanságra annotált korpuszról: a hUnCertainty korpusz magyar nyelvű Wikipédia-szócikkeit és bűnügyi híreket tartalmaz (Vincze, 2014), Vincze (2016) pedig közösségi médiából származó szövegekben foglalkozik bizonytalanság azonosításával. Az utóbbi említett két munka annotációs sémáját alkalmaztuk mi is ebben a kutatásban.

3. A StaffTalk korpusz létrehozása

A StaffTalk korpusz hétköznapi szituációkban, spontán módon létrejött nyelvi tartalmakból áll, amelyek külső hatásoknak is kitett munkahelyi környezetben keletkeztek. A korpusz ezáltal lehetővé teszi a pletyka természetének valós, munkahelyi helyzetekben történő mélyebb megértését. A korpuszt spontán nyelvi produktumok alkotják, vagyis a kutatásban résztvevők szabadon megválaszthatták beszélgetésük tárgyát, hosszát és partnereit.

A beszélgetések rögzítése egy magyarországi iskola épületében zajlott 27 munkanapon keresztül. Az adatfelvétel 2019. április 8. és május 17. között zajlott (egy kisebb megszakítással, amikor iskolai program miatt pár nap kimaradt). Az adatfelvétel során a tanári közösség által legsűrűbben használt térre, a tanári szobára fókuszáltunk. A tanári közösség azon tagjai (összesen 20 fő), akik önként vállalták a kutatásban való részvételt, egy okosórát viseltek, mellyel rögzítettük a beszélgetéseiket. Az okosóra típusa Huawei SmartWatch 1 volt.¹ A rögzítő eszköz viselő személyek a nyakukban lévő figyelmeztető felirattal jelezték a rögzítés tényét a környezetükben lévő, kutatásban részt nem vevő személyek számára, akik szintén nyilatkoztak arról, hogy hozzájárulnak a hangfelvételeken történő szerepléshez. A hangrögzítés kizárólag abban az esetben indult el, amennyiben két eszköz megfelelő közelségbe került egymással és bármikor megállítható volt. Az etikai előírásoknak megfelelően, ha valaki úgy gondolta utólag, hogy az elhangzott beszélgetést mégsem szeretné rögzíteni, akkor azt utólag is jelezhetette, természetesen ez esetben a már felvett hanganyagot töröltük. Az órák összesen 215:26:18 időtartamú hanganyagot rögzítettek.

A résztvevő tanárok többsége (70%) nő volt, átlagéletkoruk pedig 47 év (szórás: 9 év). 20 százalékuk rendelkezett a tanári munkáján felül valamilyen egyéb kiemelt beosztással: az intézményvezető, egy intézményvezető-helyettes, valamint két munkaközösség-vezető egyezett bele a részvételbe.

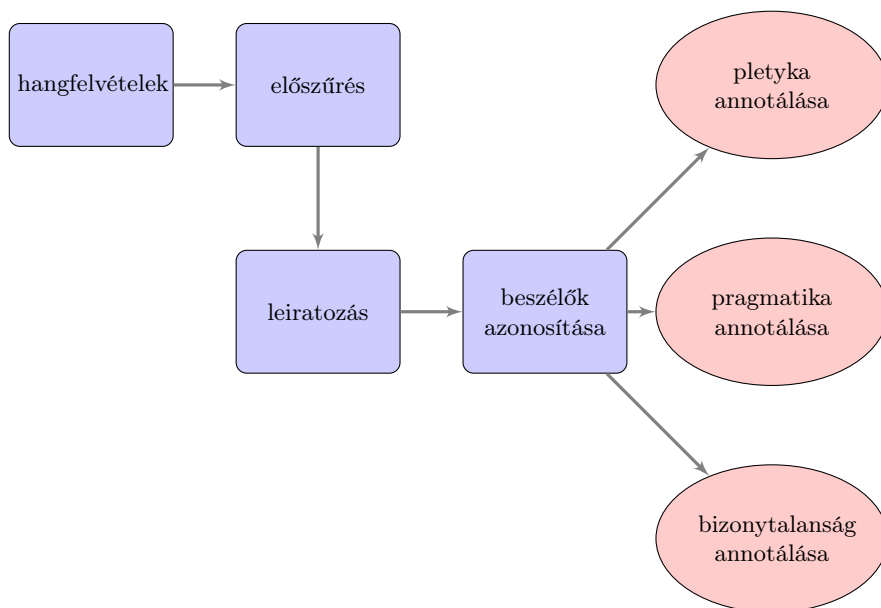
A projekt előkészítő szakaszában első lépésként a hangfájlokból kivágtuk a tíz másodpercnél hosszabb csendeket. Ennek eredményeként egy összesen 154:14:32 időtartamú hanganyag jött létre. Mielőtt a leiratozás megkezdődött, a hangfájlokat előválogattuk, amelynek során kiszűrtük a kutatás szempontjából nem releváns, adatvédelmi szempontból problémás, valamint nagyon rossz minőségű fájlokat. Amennyiben ugyanis például az okosórát viselő tanár elfelejtette az óráját kikapcsolni, az természetesen akár tanórákat, telefonbeszélgetéseket vagy diákokkal és szüleikkel történő beszélgetéseket is felvehetett, azonban ezeket kizártuk a feldolgozandó anyagok köréből. Nem foglalkoztunk az iskolai ünnepek, tanári értekezletek és a kutatókkal való egyeztetések felvételeinek a feldolgozásával sem. A csak háttérzajt tartalmazó vagy nem megfelelő hangminőségű, leiratozásra alkalmatlan felvételeket ugyancsak kivettük a végleges korpuszból. Az előválogatás után 101:07:49 időtartamú hanganyag maradt (közel 47%-a az eredeti felvételeknek), a feldolgozás során ennek leiratozása, majd annotálása történt meg.

Ami az anyagok spontaneitását illeti, bár a résztvevők tudatában voltak annak, hogy beszélgetéseikről hangfelvételek készülnek, a folyamatos és hosszabb időn át tartó rögzítésnek köszönhetően spontán beszédnek tekinthetjük a felvett beszédanyagokat.

¹ Rögzítési frekvencia: 16kHz; csatornák: 1 (mono); bitmélység: 16 bit; nyers formátum: PCM; utófeldolgozás: az emberi hangos felerősítése, valamint a halk részek, végül a teljes anyag hangosítása.

4. A korpusz feldolgozása

Ebben a részben részletesen bemutatjuk a hanganyagok feldolgozásának teljes folyamatát, melyet az 1. ábra szemléltet.



1. ábra: A munkafolyamat.

4.1. A hanganyag leiratozása

Az előzetesen kiválogatott hangfelvételekhez – a további feldolgozást és annotációt megkönnyítendő – első lépésben leiratok készültek.

Az iskolai környezetből adódóan a felvételek némelyike zajos, nehezebben érthető. Sokszor többszereplős beszélgetéseket is tartalmaz a hanganyag, ahol a szereplők gyakran egymás szavába vágtak, egyszerre beszéltek. A jelenleg a magyar nyelvhez rendelkezésre álló automatikus beszédfelismerő alkalmazások e jelenségek többségére nincsenek felkészítve, így ezek használatát elvetettük, és a kézi leiratozás mellett döntöttünk.

Ebben a fázisban tíz gépelő vett részt, akik a hallott anyagot legépeltek, időbélyegekkal, illetve különféle annotációkkal látták el. Elkülönítették egymástól az egyes beszélők megszólalásait, valamint bizonyos hanghatásokat és beszédjellemzőket (pl. suttogás, nevetés) is jelöltek. Ezeken felül az egyszerre beszélést, egymás szavába vágást is jelölték, valamint a nem és nem biztosan jól értett részeket is. Végezetül az egyes beszélgetések határait is annotálták a leiratok-

ban. A munkafolyamat során az F4 szoftvert használták². Egy órányi hanganyag leiratozása átlagosan 14 órájába telt a gépelőknek.

A leiratozás minőségének biztosítása érdekében rendszeres időközönként bizonyos hangfájlokat (összesen a korpusz közel 10%-át) több gépelővel is leírtattunk, majd az így kapott szövegváltozatokat automatikusan összevetettük egymással. Vizsgáltuk a szókincs egyezését, valamint a szövegben elhelyezett annotáció minőségét és az egyes tagek mennyiségét. Ha az egyezés mértéke nem felelt meg az előzetes elvárásoknak (60%), akkor a fájlokat további javításra visszaadtuk az adott leiratozónak, ez körülbelül az ellenőrzött fájlok 15%-át érintette.

4.2. A beszélők azonosítása

Fő kutatási kérdéseink egyike volt, hogy egy zárt csoporton belül hogyan befolyásolja a kommunikáció, különösen a pletyka a közösség működését. Ennek vizsgálatához elengedhetetlen, hogy rendelkezésünkre álljon az adott beszélgetésben részt vevő személyek kiléte is, azaz tudjuk, ki, mikor, kivel és miről beszélgetett. Elengedhetetlen volt tehát az egyes diskurzusokban részt vevő személyek név szerinti azonosítása is.

A kutatás 20 résztvevőjéről rendelkezésre álltak hangminták is, illetve ismert volt az az információ is, hogy melyik órát ki viseli. Ugyanakkor a résztvevők beszélgethettek olyan személyekkel is, akik nem vettek részt az adatfelvételben (pl. diákok, szülők), így a hangfelvételeken megszólaló személyek száma meghaladja a 30-40-et is. Mivel előzetes tapasztalataink azt mutatták, hogy a gépelést jelentősen lelassítja, ha a leiratozónak kell egyúttal azonosítani is a beszélőket, külön munkafázisba szerveztük ezt a feladatot: két erre szakosodott nyelvész (kutatói támogatással) végezte a már elkészített leiratok alapján az egyes megszólalók azonosítását. A kutatásban részt nem vevők hangját külön azonosítóval („külső személy”) láttuk el.

A fentebbi munkaszervezési megoldással nagymértékben sikerült meggyorsítani a leiratozás folyamatát, valamint hatékonyabbá tenni a beszélők azonosítását.

E fázist követően a létrejött szövegfájlokat három különálló fázisban annotáltuk, amelyhez az MMAX2 eszközt (Müller és Strube, 2006) használtuk. Az annotátorok időnként ugyanazokat a fájlokat annotálták egymástól függetlenül, így munkájuk minőségét össze tudtuk vetni. A minőségbiztosítás fájljainak mennyiségét úgy határoztuk meg, hogy kitegye a teljes korpusz 10%-át, és ezáltal megfeleljen a nemzetközi sztenderdnek. Annak céljából pedig, hogy az ellenőrzés folyamatos lehessen, ezeket a fájlokat a munka teljes hosszában arányosan osztottuk ki. Amennyiben az annotátorok közti egyetértés nem érte el a kívánt szintet, a fájlokat utólag ki kellett javítaniuk, erre azonban az eseteknek csupán a töredékében (körülbelül 15%-ában) volt szükség, és az annotátoroknak minden esetben sikerült a problémás fájlokat megfelelően korrigálni.

Az egyes fázisok részleteit az alábbiakban ismertetjük.

² <https://www.audiotranskription.de/english/f4>

4.3. Pletyka

A *pletyka* annak hagyományos definíciója szerint olyan, jelen nem lévő személyről vagy személyekről folytatott értékelő tartalmat hordozó beszélgetéseket takar, amelyekben azt értékelést megfogalmazó személy és legalább egy hallgató jelen van (Emler, 1994; Grosser és mtsai, 2012). A *pletyka* fogalma gyakran negatív konnotációt hordoz, azonban fontos szerepet tölthet be az információáramlásban, a személyek közti kapcsolatok megerősítésében, a csoportnormák fenntartásában és betartatásában vagy szelepként szolgálhat a felgyülemlett negatív érzelmek „kiengedésében” (Grosser és mtsai, 2012).

A jelen tanulmányban bemutatott korpusz építése során egyéb, személyekre vagy a köztük lévő kapcsolatra irányuló beszéd tartalmakat is vizsgálunk. A személyekre vonatkozó beszéd tartalmak annotálását két fő dimenzió mentén végeztük, a pletyka célszemélye vagy célszemélyei tekintetében. Csoporton belüli pletykának tekintettük az annotálás során, amennyiben a pletyka célszemélye a munkatársi csoporthoz tartozó személy, míg csoporton kívüli pletykának, amennyiben a célszemély nem a munkatársi csoporthoz tartozik. A munkatársi csoporthoz tartozó és nem tartozó személyekről szóló beszéd tartalom megkülönböztetése azért releváns, mert eltérő funkciót tölthetnek a szervezeten belüli kommunikációban.

Jelenlévő célszemélyre vonatkozó beszéd tartalomnak tekintettük, amennyiben az összes, a pletykában említett célszemély jelen volt a beszélgetésben, míg jelen nem lévő célszemélyre vonatkozó beszéd tartalomnak, amennyiben az említett személy vagy személyek nem voltak jelen a beszélgetésben. A munkatársi csoporthoz tartozó és nem tartozó, valamint a jelenlévő és jelen nem lévő célszemélyekre vonatkozó pletyka vegyesen is előfordulhat egy pletykán belül (főként személyek közti relációra vonatkozó tartalmak esetén), így ezeket is megkülönböztettük. A két dimenzió mentén való besoroláson felül pedig további ismérvek jelölését is kértük az annotátoroktól a személyekre vonatkozó beszéd tartalmakra nézve, az alábbiaknak megfelelően:

- típus: csoporton belüli vagy kívüli személyekre, illetve jelen lévő vagy jelen nem lévő személyekre irányul a pletyka
- polaritás: pozitív, negatív vagy semleges a szövegtartalom
- forrás: a pletyka közlöje
- célpont: az a személy, akiről szól a pletyka
- reláció: két vagy több személy relációjáról szól-e a pletyka
- normativitás: normatív magatartással függ-e össze a pletyka

Jelölték ezen felül, amennyiben egy pletykára a pletykát megerősítő, továbbvivő, illetve azt hárító reakció érkezett. Amennyiben a pletykára érkező reakció egyben új pletykát is tartalmazott, akkor az adott szövegrészt pletyaként és reakcióként is annotálták.

Több pletyka összefűzhető volt egy láncba, amennyiben a pletyka ugyanazon témában ugyanazon célszemélyre vonatkozott.

A fentebb bemutatottakon túl a korpusz egyéb információk kinyerését is lehetővé teszi, amelyek annotálására nem volt szükség. Így például, mivel a hangok

azonosítása is megtörtént, akár a munkahelyi hierarchia, illetve az életkor és a pletyka összefüggései is vizsgálhatóak lesznek a korpusz segítségével a jövőben.

4.4. Pragmatika

A közösségen belüli kommunikáció vizsgálatának egyik fontos vetülete, hogy milyen beszédaktusokat és udvariassági stratégiákat használnak egymás között az egyes közösségi tagok. Ennek vizsgálatához részletes pragmatikai annotációval láttuk el a leiratokat, az alábbi annotációs sémát alkalmazva.

- Beszédaktusok:
 - ígéret / ajánlat
 - figyelmeztetés / fenyegetés
 - kérés / parancs / kívánság
 - panasz / vád / kritika / sértés
 - dicséret / bók
 - bocsánatkérés
 - köszönetnyilvánítás
- Reakciók:
 - elfogadás / egyetértés
 - visszautasítás / egyet nem értés
 - háritás (nem egyértelmű elfogadás vagy visszautasítás)
- Irónia:
 - irónia
 - antiirónia
- Interakciós elemek:
 - figyelem felhívása
 - üdvözlés / elköszönés

A pragmatikaihoz annotációt két nyelvész készíttette a már említett MMAX2 szoftver (Müller és Strube, 2006) segítségével. A pragmatikai (és bizonytalansági) annotációkat részletesen taglaljuk egy másik, az MSZNY2021 konferencián megjelent cikkben (Vincze és mtsai, 2021).

4.5. Bizonytalanság

Harmadik annotációs szintként az annotátorok megjelölték a nyelvi bizonytalanságra utaló szavakat a korpuszban. Úgy gondoljuk, hogy a bizonytalanság annotálása összeköthető mindkét másik annotációs szinttel: egyfelől feltételezzük, hogy a pletyka közlője bizonytalanságot hordozó nyelvi elemeket is beleszőhet a mondandójába, ami a pletykában az egyes típusok gyakoriságát illetően egyfajta, eddig ismeretlen mintázatot mutathat. Másfelől bizonyos beszédaktusok és a bizonytalanság kifejezőeszközei sokszor egybeesnek (például kérésekben gyakran szerepel feltételes módú ige), így érdemes összevetni ezen nyelvi elemek többféle szerepét ugyanabban a nyelvi adatbázisban.

A bizonytalanság annotálásakor követtük a már korábban létrehozott magyar nyelvű bizonytalansági korpuszok kategorizálását (Vincze, 2014, 2016), amelyet az alábbiakban foglalkunk össze:

- Szemantikus bizonytalanság:
 - episztemikus
 - doxasztikus
 - feltételes
 - vizsgálat
- Diskurzusszintű bizonytalanság:
 - weasel: bizonytalan információforrás vagy szereplő a cselekvésben
 - hedge: mennyiségek vagy minőségek homályos jelölése
 - peacock: bizonyít(hat)atlan állítás vagy túlzás

A bizonytalanság annotálását – a pragmatikaihoz hasonlóan – két nyelvész végezte az MMAX2 szoftver (Müller és Strube, 2006) segítségével. Ahogy már fentebb említettük, a bizonytalansági annotációkat is részletesen elemezzük egy másik, az MSZNY2021 konferencián megjelent cikkben (Vincze és mtsai, 2021).

5. Statisztikai adatok

A korpuszban található annotációk mennyiségi megoszlása az 1. táblázatban látható.

Annotációs szint	Hanganyag időtartama	Mondatszám	Tokenszám	Annotált egységek száma
Pletyka	102:42:30	124 836	1 461 769	44 165
Pragmatika	102:42:30	124 836	1 461 769	26 463
Bizonytalanság	102:42:30	124 836	1 461 769	28 340

1. táblázat. A korpusz adatai.

6. A korpusz felhasználhatósága

A StaffTalk korpusz – kézi leiratozása és részletes annotációja, továbbá keletkezési körülményei miatt – egyaránt hasznos lehet mind a beszédtechnológusoknak, mind számítógépes nyelvészeknek, elméleti és alkalmazott nyelvészeknek, valamint a társadalomtudósoknak.

A korpuszépítés befejezését követően a korpuszt adatvédelmi okok miatt anonimizáljuk. Az anonimizálást követően a korpuszt bárki számára kutatási és oktatási célra elérhetővé tesszük.

7. Összegzés

A dolgozatban bemutattunk a StaffTalk korpuszt, amely a magyar nyelvű spontán diskurzust reprezentálja, a hangzó szövegek legépelt és annotált változatával együtt. A legépelt hanganyagokat három szinten annotáltuk: a beszélgetésekben

megjelenő pletykát, a beszédaktusokat és egyéb pragmatikai jegyeket, valamint a bizonytalanságra utaló szavakat egyaránt megjelöltük. A tanulmány célja az volt, hogy részletes információval szolgáljon a korpusz készítésének céljáról, eszközeiről és módjáról, valamint ismertesse annak alapvető statisztikai adatait.

A korpuszban foglalt szöveganyag, valamint a feldolgozás módja teret nyit számos olyan vizsgálat elvégzésére a jövőben, amely a spontán beszéd természetét kívánja kutatni, válaszokat adva ezzel bizonyos, a humán kommunikáció, illetve interakció természetét érintő kérdésekre. Mind a pletykát érintő, mind az udvariasság és bizonytalanság témakörébe tartozó kutatási kérdéseinket szeretnénk behatóan vizsgálni és tárgyalni a jövőben a korpusz segítségével.

A hangrögzítést kiegészítette egy napi szintű kérdőíves felmérés, mely a kutatásban résztvevők informális és formális kapcsolataira, munkahelyi elégedettségére és közérzetére vonatkozó kérdéseket tartalmazott. A kiegészítő adatok lehetővé teszik a pletyka más hálózatokkal való összefüggésben történő vizsgálatát is.

Tervezzük a korpusz nyilvánossá tételét a jövőben a kutatók számára, az adatok anonimizálását követően.

Köszönetnyilvánítás

A kutatást az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében támogatta az ERC_CoG_2014_648693 sz. szerződésben, a kutatás vezetője Takács Károly.

Szabó Martina Katalin kutatásait részben a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal – NKFIH OTKA posztdoktori kiválósági programja (NKFI-azonosító: 132312) támogatta.

Szeretnénk köszönetet mondani a korpusz leiratozásában és annotálásában tevékenyen részt vállaló kollégáinknak kitartó és lelkes munkájukért.

Hivatkozások

- Babarczy, A.: Analógikus általánosítási folyamatok a gyereknyelvben= analogical generalisation processes in language acquisition. OTKA Kutatási Jelentések| OTKA Research Reports (2009)
- Bodó, Cs., Kocsis, Zs., Vargha, F.: A Budapesti Egyetemi Kollégiumi Korpusz. Elméleti és módszertani kérdések. In: Benő, A., Fazakas, N. (szerk.) Élőnyelvi kutatások és a dialektológia: Válogatás a 19. Élőnyelvi Konferencia - Marosvásárhely, 2016. szeptember 7-9. - előadásából. pp. 169–177 (2017)
- Calhoun, S., Carletta, J., Brenier, J.M., Mayo, N., Jurafsky, D., Steedman, M., Beaver, D.: The next-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language resources and evaluation* 44(4), 387–419 (2010)
- Crowdy, S.: Spoken corpus design. *Literary and Linguistic Computing* 8(4), 259–265 (1993)

- Deme, A., Markó, A.: Lengthenings and filled pauses in Hungarian adults' and children's speech. In: Proceedings of DiSS 2013, The 6th Workshop on Disfluency in Spontaneous Speech. TMH-QPSR 54:1. vol. 54, pp. 21–24. KTH Royal Institute of Technology (2013)
- Dunbar, R.I.: Grooming, Gossip and the Evolution of Language. Harvard University Press, Cambridge, MA (1996)
- Dunbar, R.I.: Gossip in evolutionary perspective. *Review of General Psychology* 8(2), 100–110 (2004)
- Eckhaus, E., Ben-Hador, B.: Gossip and gender differences: a content analysis approach. *Journal of Gender Studies* 28(1), 97–108 (2019)
- Emler, N.: Gossip, reputation and social adaptation. In: Goodman, R.F., Ben-Ze'ev, A. (szerk.) *Good gossip*. University Press of Kansas, Lawrence (1994)
- Foster, E.K.: Research on gossip: Taxonomy, methods, and future directions. *Review of general psychology* 8(2), 78–99 (2004)
- Galántai, J., Pápay, B., Kubik, B.G., Szabó, M.K., Takács, K.: A pletyka a társas rend szolgálatában-az informális kommunikáció struktúrájának mélyebb megértéséért a computational social science eszközeivel. *Magyar Tudomány* 179(7), 964–976 (2018)
- Gósy, M.: BEA–A multifunctional Hungarian spoken language database. *Phonetician* 105, 50–61 (2013)
- Grosser, T., Kidwell, V., Labianca, G.J.: Hearing it through the grapevine: Positive and negative workplace gossip. *Organizational Dynamics* 41, 52–61 (2012)
- Gulyás, A., Galántai, J., Szabó, M.K., Szebeni, Z.: A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 317–330. Szegedi Tudományegyetem, Szeged (2018)
- Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The ATIS spoken language systems pilot corpus. In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24–27, 1990* (1990)
- Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Van Ess-Dykema, C.: Automatic detection of discourse structure for speech recognition and understanding. In: *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. pp. 88–95. IEEE (1997)
- Kane, J., Pápay, K., Hunyadi, L., Gobl, C.: On the Use of Creak in Hungarian Spontaneous Speech. In: *ICPhS*. pp. 1014–1017 (2011)
- Kim, J.D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9(Suppl 10) (2008), <http://www.biomedcentral.com/1471-2105/9/10>
- Kugler, N.: Megfigyelés és következtetés a nyelvi tevékenységben. No. 179, Tinta Könyvkiadó (2015)
- Leech, G., McEnery, T., Weisser, M.: Spaac speech-act annotation scheme. University of Lancaster (2003)
- Lennes, M., és mtsai: Segmental features in spontaneous and read-aloud finnish. *Phonetics of Russian and Finnish general description of phonetic systems: experimental studies on spontaneous and read-aloud speech* (2009)

- Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous Speech Corpus of Japanese. In: LREC. pp. 947–9520. Citeseer (2000)
- McEnery, T.: Corpus linguistics, vol. 978019. Oxford University Press Inc (2012)
- Mengusoglu, E., Deroo, O.: Turkish lvcslr: Database preparation and language modeling for an agglutinative language. In: IEEE International Conference on Acoustics Speech And Signal Processing. vol. 6, pp. 4018–4018. IEEE; 1999 (2001)
- Mitra, T., Gilbert, E.: Have you heard?: How gossip flows through workplace email. In: ICWSM (2012)
- Müller, C., Strube, M.: Multi-level annotation of linguistic data with MMAX2. In: Braun, S., Kohn, K., Mukherjee, J. (szerk.) Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods, pp. 197–214. Peter Lang, Frankfurt a.M., Germany (2006)
- Neuberger, T., Gyarmathy, D., Grácsi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative Hungarian language. In: International Conference on Text, Speech, and Dialogue. pp. 424–431. Springer (2014)
- Oostdijk, N.: The spoken dutch corpus. overview and first evaluation. In: LREC. pp. 887–894. Athens, Greece (2000)
- Pápay, K., Szeghalmy, S., Szekrényes, I.: Hucomtech multimodal corpus annotation. Argumentum 7, 330–347 (2011)
- Reichel, U.D., Mády, K.: Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech. In: Wagner, P. (szerk.) Elektronische Sprachsignalverarbeitung 2013. pp. 223–230. TUDpress, Dresden (2013), <http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-18043-4>
- Robbins, M.L., Karan, A.: Who gossips and how in everyday life? Social Psychological and Personality Science 11(2), 185–195 (2020)
- Saurí, R., Pustejovsky, J.: FactBank: a corpus annotated with event factuality. Language Resources and Evaluation 43, 227–268 (2009), <http://dx.doi.org/10.1007/s10579-009-9089-9>
- Seppänen, T., Toivanen, J., Väyrynen, E.: MediaTeam speech corpus: a first large Finnish emotional speech database. In: Proceedings of the Proceedings of XV International Conference of Phonetic Science. pp. 2469–2472. Citeseer (2003)
- Szabó, M.K., Galántai, J.: Egy magyar nyelvű spontán beszélt nyelvi korpusz (HuTongue) létrehozásának tapasztalatai. In: XXVI. MANYE Kongresszus konferenciakötete. Pécs (2017)
- Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., Gurevych, I.: Cross-genre and cross-domain detection of semantic uncertainty. Computational Linguistics 38, 335–367 (June 2012)
- Van Bael, C., Baayen, R.H., Strik, H.: Segment deletion in spontaneous speech: a corpus study using mixed effects models with crossed random effects. In: INTERSPEECH. pp. 2741–2744 (2007)
- Váradi, T.: A budapesti szociolingvisztikai interjú. In: Kiefer, F., Siptár, P. (szerk.) A magyar nyelv kézikönyve, pp. 339–359. Akadémiai Könyvkiadó, Budapest (2003)

- Vicsi, K., Tóth, L., Kocsor, A., Csirik, J.: MTBA—a Hungarian telephone speech database. *Híradástechnika*, LVII 8 (2002)
- Vincze, V.: Uncertainty detection in Hungarian texts. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. pp. 1844–1853. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (Aug 2014), <https://www.aclweb.org/anthology/C14-1174>
- Vincze, V.: Detecting uncertainty cues in Hungarian social media texts. In: *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*. pp. 11–21. The COLING 2016 Organizing Committee, Osaka, Japan (Dec 2016), <https://www.aclweb.org/anthology/W16-5002>
- Vincze, V., Üveges, I., Szabó, M.K.: Magyar nyelvű spontán beszéd szemantikai–pragmatikai sajátosságainak elemzése nagy méretű korpusz (StaffTalk) alapján. In: *XVII. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Szeged (2021)
- Zhu, X., Penn, G.: Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. pp. 197–200 (2006)

Automatikus írásjel-visszaállítás és Nagybetűsítés statikus korpuszon transzformer modellen alapuló neurális gépi fordítással

Yang Zijian Győző

¹MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.
yang.zijian.gyozo@itk.ppke.hu

Kivonat Cikkemben egy írásjelvisszaállító és nagybetűsítő programot mutatok be, amelyet a jelenkori „state-of-the-art” transzformer modellen alapuló neurális gépi fordító rendszerrel tanítottam be. A mobil eszközökön történő üzenetírás elterjedésével és a minél gyorsabb szövegbevitelre való törekvéssel tömeges jelenséggé vált a hibás szövegek írása. Ennek egyik következménye, hogy a interneten elérhető – főleg a szociális médiából származó – korpuszok egy része hibás. Többek között írásjelek hiányoznak, vagy végig kisbetűvel írnak. Az így létrejött korpuszok nem alkalmasak különböző kutatásokhoz, csak tisztítás után. A tisztítás folyamata időigényes, ezért igény van különböző korpusztisztító módszerekre. Az általam létrehozott rendszer, annak ellenére, hogy semmilyen morfológiai és szintaktikai elemzőt nem használ, közel 81%-os f-mértékkel tudja helyesen visszaállítani az alapírásjeleket és elvégezni a nagybetűsítést magyar nyelv esetében.

Kulcsszavak: írásjel-visszaállítás, nagybetűsítés, neurális háló-alapú gépi fordítás, NMT, transzformer modell

1. Bevezetés

Napjainkban a számítógépes nyelvészek számára nagy lehetőséget nyújtanak az interneten elérhető nagy mennyiségű szövegek. Számos részterületen használjuk a weboldalakról összegyűjtött korpuszokat, mint például a gépi fordítás, a szövegkivonatolás vagy az érzelemdetektálás. Ezekhez a feladatokhoz viszont nélkülözhetetlen, hogy a vizsgált szöveg a lehető legjobb minőségű legyen.

A mobil eszközökön írt szövegek és üzenetek esetében tömegjelenséggé vált az ékezetes betűk és írásjelek elhagyása. Ennek következményeként léteznek olyan korpuszok is, amelyek egy része ékezet- és írásjelmentes, vagy kisbetűvel írták. Így nem működnek rajtuk a természetes szövegen betanított szövegfeldolgozó modellek.

Magyar nyelvre Dömötör és Yang (2018) végeztek kutatást különböző korpuszokban a nem sztenderd hibák előfordulására. Első sorban beszélt nyelvi és személyes alkörpuszokból indultak ki, hiszen ott lehet nagyobb mennyiségben

hibás szöveget találni. Kimutatták, hogy a nem sztenderd hibák közel 30%-át kiugróan az írásjelek és nagybetűk elhagyása teszi ki.

Az elmúlt években a neurálishálózat-alapú módszerek eredményei túlszárnyalták az addigi legjobb rendszereket. Ez a nyelvtchnológia területén is megmutatkozik, ezért célom az volt, hogy megvizsgáljam az írásjel- és nagybetű-visszaállítás problémáját a jelenlegi „state-of-the-art” NMT-alapú rendszerrel.

2. Kapcsolódó irodalom

Beszédtechnológia terén az írásjel-visszaállítás egy fontos feladat. Nehézsége abban rejlik, hogy a szöveg dinamikusan változik és mindig az adott környezethez igazodva kell visszaállítani az írásjelet.

Öktem és mtsai (2017) az automatikus beszédfelismerés feladatában az RNN neurális hálózat segítségével állítják vissza az írásjeleket. Transformers modell alapú írásjel- és nagybetű-visszaállítást Vāravcs és Salimbajevcs (2018) végezték lett és angol nyelvre. Kutatásukban a neurális gépi fordítást használták rendszerük betanítására. Nguyen és mtsai (2019) kutatásukban transformer modellel az írásjel- és nagybetű-visszaállítás mellett főnévi csoportok felismerés feladatát is belevették a tanításba. Alam és mtsai (2020) az angol mellett a kevés tanítóanyaggal rendelkező bengáli nyelvre tettek kísérletet az írásjelek visszaállítására. Mivel kevés a nyersanyag, ezért a különböző BERT (Devlin és mtsai, 2019) alapú modellek finomhangolásával tanítottak be modelleket, amelyekkel a problémát megoldják.

Magyar nyelvre Tündik és mtsai (2018) az RNN hálózat segítségével állítják vissza az írásjeleket.

Kutatásom nem a dinamikusan változó szövegekre koncentrál, hanem a statikus korpuszokra. Módszerem első sorban korpusztisztításra alkalmas.

3. A korpusz és a fordító rendszer

A korpusz-alapú gépi fordító rendszer lényege, hogy transzformációt képez tet-szőleges forrás- és célnyelvi mondatok között, ahol a rendszer betanításához nem kell más, mint egy kétnyelvű párhuzamos korpusz. Az írásjelek és nagybetűk helyreállítására a gépi fordítás módszereit választottam, mivel az írásjellel ellátott nagybetűkkel rendelkező mondatok grammatikailag, szókinccsal és szó szerkezetileg nagyon hasonlóak az írásjel és nagybetű nélküli párjukhoz.

A neurális hálózat tanításához nagy mennyiségű tanítóanyagra van szükség, melynek előállítás a jelen feladathoz igen könnyű. A tanítóanyag létrehozásához annyit kellett tenni, hogy egy egynyelvű korpusz írásjeleit eltávolítottam és a szöveget kisbetűsíttem.

A korpusz létrehozásához az online elérhető Open Subtitles¹ nevű angol-magyar párhuzamos korpuszának magyar oldali szövegét használtam. A korpusz

¹ <http://opus.nlpl.eu/OpenSubtitles-v2018.php>

TV és mozi filmekre létrehozott feliratokból áll. Ennek megfelelően főleg rövidebb, informális mondatokat tartalmaz. A gépi fordító rendszer célnyelvi korpuszának előállításához a mondatokban az írásjeleket kitöröltem, majd a szöveget kisbetűsítettem.

A korpusz megközelítőleg 29 millió szegmensből áll, melyből 5000 mondatot validációs és 3000 mondatot tesztelési célra elkülönítettem. A korpusz az egyik legnagyobb szabadon hozzáférhető párhuzamos tanítóanyagának számít, ellenben mérete elmarad az egy nyelvű tanítóanyagokétól. Választásom azért esett erre az adathalmazra, mert több párhuzamos kutatásom során is használtam, és néhány koprusztisztító lépést már előzetesen eszközöltünk rajta. Kivettem azokat a mondatokat, amelyek speciális karaktereket (pl. kínai, japán, cirill stb.) tartalmaztak, valamint a teszt halmaz mondatait kézzel kijavítottam. Mérete elégséges a neurális hálózatok helyes betanítására, valamint a tanítási idő is viszonylag kezelhető marad (1-2 nap). Végül utolsó szempont, hogy a feliratok gyakran hasonló mondat szerkezetűek, mint a beszélt nyelvi mondatok, amelyekben a legtöbb nem sztenderd hiba található.

A 2010-es évek első felére a statisztikai gépi fordítórendszerek elérték teljesítmképességük határát. Az alapjait képező módszert és a létrehozott keretrendszereket a kutatók nagyon sok befektetett munkája ellenére lényegében nem sikerült tovább javítani. Az áttörést (Bahdanau és mtsai, 2015) rendszere hozta el, ami egy figyelmi modellel támogatott enkóder-dekóder architektúrájú NMT rendszer volt. A modell lényege, hogy kettéválasztja a fordítás folyamatát két elkülöníthető részre. A kódolás során lényegében egy RNN-alapú seq2seq modellt hoz létre, tehát a szóbeágyazási modellhez hasonlóan a fordítandó modellekből egy n -dimenziós vektort készít. 1. ábrán ez a vektor felel meg az ábra közepén látható piros/sötét node-nak. A második fázis a dekódolás, ahol a mondatvektorból generálja ki a cél nyelvi mondatot egy RNN réteg segítségével.

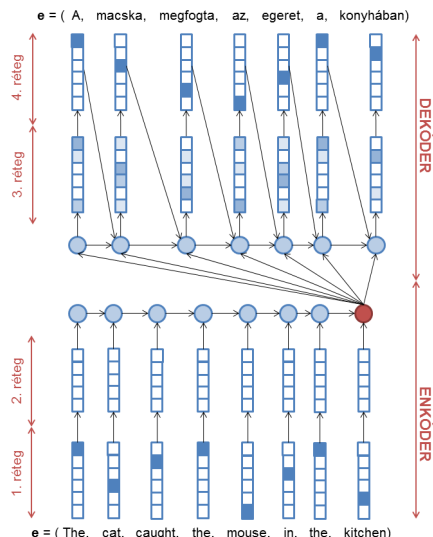
Innentől számítva az NMT rendszerek átvették a vezető szerepet az SMT-től. 2017-ben a Google cég munkatársai (Vaswani és mtsai, 2017) publikálták és szabadon hozzáférhetővé tették az úgynevezett multi-attention réteggel támogatott NMT rendszerüket. Ezt a szakirodalomban transzformer-alapú architektúrának nevezik. A módszer lényege, hogy az eddigi egy helyett több figyelmi réteget helyeztek el a rendszerben, ami segítségével nagymértékben nőtt a többértelmű szavak fordításának minősége.

Munkám során a Marian NMT (Junczys-Dowmunt és mtsai, 2018) nevű keretrendszert használtam, ami egy `c++` nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimalis implementációjának köszönhetően² az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

Manapság a neurális hálózat alapú modellek tanításához részszó (subword) tokenizálót (Sennrich és mtsai, 2015) használnak, hogy csökkentsék a szótárak méretét, és közben kezeljék az ismeretlen szavak problémáját.

A BPE (Byte Pair Encoding) egy adattömörítő eljárás, ahol a leggyakoribb bájt párokat egy olyan bájjal helyettesítjük, amely nem szerepel magában az

² <https://marian-nmt.github.io/>



1. ábra: Enkóder-dekóder architektúra vázlatos rajza

adatban. Az eljárás a korpuszon először egy karakteralapú szótárat hoz létre, ahol minden szót karakterek sorozataként ábrázol. Ezután gyakoriság alapján a gyakori karaktersorozatokat önálló tokenekként kezeli. Ezzel az adat tömörítése mellett az ismeretlen szavak kezelését is megoldja, hiszen a részszavakból előállítható egy olyan összetétel, amely nem szerepelt eredetileg a korpuszban.

Ezt a módszert fejlesztették tovább (Kudo és Richardson, 2018). Az általuk létrehozott Sentence Piece nevű eszköz egy felügyelet nélküli szöveg tokenizáló és detokenizáló, melyet elsősorban a neurálishálózat-alapú gépi tanulási feladatokhoz fejlesztettek ki. Implementálva van benne a BPE metrika, ami egy unigram nyelvmodellel (Kudo, 2018) van súlyozva. Használatával elhagyhatók a költséges nyelvspecifikus előfeldolgozási lépések, mint például a tokenizálás vagy a kisbetűsítés. A módszer lényege, hogy a természetes szöveget úgy alakítja át, hogy abban a különböző „szavak” száma korlátos legyen, valamint az így létrejött tanítóanyagban nem lesznek ismeretlen szavak. Ennek köszönhetően a neurális hálózatok paraméterszáma nagymértékben csökkenthető.

4. Kísérletek

Először megszámloltam (lásd 1. táblázat), hogy hányszor szerepelnek a számomra releváns esetek, vagyis az írásjelek és a nagybetűs szavak.

Az NMT tanításához a Marian neurális gépi fordítórendszert használtam. A rendszer fontos jellemzője a SPM technológia. A rendszerem tanításához az alábbi paramétereket használtam:

	Tanító	Valid	Teszt
nagybetűs szavak	19 586 281	53 237	31 558
"."	8 025 537	21 782	12 909
","	13 043 393	35 241	20 886
"?"	257 005	712	409
"!"	115 393	306	181
","	53 674	149	90
":	753 380	2046	1 270
"," (magyar kezdő (alsó) idézőjel)	834 363	2 183	1 286
"" (magyar záró (felső) idézőjel)	820 681	2 142	1 317
"-" (kötőjel)	372 8116	10 155	5 939
"_" (nagykötőjel)	1 026 991	2 752	1 812

1. táblázat. Írásjelek és nagybetűs szavak előfordulásai a korpuszokban

- Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
- Transformer modell: enkóder és dekóder rétegeinek száma: 6; transformer-dropout: 0,1;
- learning rate: 0,0003; lr-warmup: 16000; lr-decay-inv-sqrt: 16000;
- optimizer-params: 0,9 0,98 1e-09; beam-size: 6; normalize: 0,6
- label-smoothing: 0,1; exponential-smoothing

Kétféle modellt készítettem:

- alap: kisbetűsítés és alapírásjelek elhagyása:
 - ".": pont
 - ",": vessző
 - "?": kérdőjel
 - "!": felkiáltójel
- bővített: kisbetűsítés és bővített írásjelek elhagyása:
 - az alapírásjelek
 - ";": pontosvessző
 - ":": kettőspont
 - ",": (magyar kezdő (alsó) idézőjel)
 - """: (magyar záró (felső) idézőjel)
 - "-": (kötőjel)
 - "_": (nagykötőjel)

5. Eredmények

Kutatásom során megmértem a gép által adott szóalapú eredmény pontosságát (precision), fedését (recall) és az F-mértékét. Mivel a gépi fordítás során az eredetileg helyes szavak is megváltozhatnak, az összes szóra szükséges megvizsgálni a fordítás pontosságát (ALL). Végeztem külön kiértékelést csak azokra a szavakra, amelyek a kutatásom számára relevánsak (REL), beleértve az írásjelekkel rendelkező és nagybetűs szavakat egyaránt.

Emellett külön megmértem azt, hogy a modelljeim az írásjeleket (alap és bővített), valamint a nagybetűs szavakat milyen mértékben tudták visszaállítani.

A 2. táblázat eredményei alapján láthatjuk, hogy az általam létrehozott rendszer teljesítménye, amely transzformer modellt és Sentence Piece tokenizálót használ, az összes szóra nézve meghaladja a 92%-os F-mértéket. Az alapírásjelek és nagybetűk visszaállítását pedig 81%-os pontossággal tudja a rendszerem elvégezni. Az eredmények azt mutatják, hogy a nagybetűsítés feladatát pontosabban végzi, mint az írásjel-visszaállítást.

	Pontosság	Fedés	F-mérték
alap + nagybetű (ALL)	93,28%	91,43%	92,34%
alap + nagybetű (REL)	82,88%	79,65%	81,23%
bővített + nagybetű (ALL)	91,09%	90,31%	90,70%
bővített + nagybetű (REL)	78,72%	76,16%	77,42%
nagybetű	85,88%	83,33%	84,59%
alapírásjelek	79,60%	75,38%	77,43%
bővített írásjelek	73,89%	70,17%	71,98%

2. táblázat. Írásjel-visszaállítás és nagybetűsítés eredményei

	Pontosság	Fedés	F-mérték
alap "."	74,43%	73,09%	73,75%
alap ", "	82,89%	77,95%	80,35%
alap "? "	67,15%	38,49%	48,93%
alap "! "	52,17%	9,75%	16,43%
bővített "."	70,88%	73,57%	72,19%
bővített ", "	78,08%	78,77%	78,42%
bővített "? "	66,66%	36,67%	47,31%
bővített "! "	58,69%	14,91%	23,78%

3. táblázat. Alap írásjelek visszaállításának részletes eredményei

Végül kiértékeltem a két modell alapírásjeleinek visszaállításának teljesítményét külön-külön (lásd 3. táblázat). A 3. táblázat eredményei alapján a kérdőjelek és a felkiáltójelek értékei alacsonyak. Ez annak tulajdonítható, hogy a tesztanyagban ezek az írásjelek elég kevészer szerepelnek (lásd 1. táblázat). Legjobban a vesszőket tudja a rendszer visszaállítani, közel 80%-os F-mértékkel.

6. Összegzés

A kutatásommal létrehoztam egy írásjel- és nagybetű-visszaállító rendszert. A rendszer tanításához egy neruálishálózat-alapú gépi fordítórendszert használtam,

amely transzformer modellt és Sentence Piece tokenizálót használ. A rendszerem 81%-os F-mértékkel tudja helyesen visszaállítani az alapírásjeleket és a nagybetűket.

Továbblépési lehetőségként szeretném a modelleket a beszédtechnológia feladataira hangolni, valamint kipróbálni az új BERT alapú modellek teljesítményét is.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1-NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

- Alam, T., Khan, A., Alam, F.: Punctuation restoration using transformer models for resource-rich and -poor languages. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). pp. 132–142. Association for Computational Linguistics, Online (Nov 2020)
- Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (szerk.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), <http://arxiv.org/abs/1409.0473>
- Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Mázler, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019), <http://www.aclweb.org/anthology/W19-5301>
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
- Dömötör, A., Yang, Z.G.: Így írtok ti: nem sztenderd szövegek hibatípusainak detektálása gépi tanulásos módszerrel. XIV. Magyar Számítógépes Nyelvészeti Konferencia pp. 305–316 (2018)
- Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)

- Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 66–75. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
- Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018), <https://www.aclweb.org/anthology/D18-2012>
- Nguyen, B., Nguyen, V., Nguyen, H., Pham, P., Nguyen, T.L., Do, T., Luong, C.: Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging. pp. 1–5 (10 2019)
- Öktem, A., Farrús, M., Wanner, L.: Attentional parallel rnns for generating punctuation in transcribed speech. In: Camelin, N., Estève, Y., Martín-Vide, C. (szerk.) Statistical Language and Speech Processing. pp. 131–142. Springer International Publishing, Cham (2017)
- Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CoRR abs/1508.07909 (2015), <http://arxiv.org/abs/1508.07909>
- Tündik, M.Á., Tarján, B., Szaszák, Gy.: Televíziós feliratok írásjeleinek visszaállítása rekurrens neurális hálózatokkal. XIV. Magyar Számítógépes Nyelvészeti Konferencia pp. 183–195 (2018)
- Vāravs, A., Salimbajevs, A.: Restoring punctuation and capitalization using transformer models. In: Dutoit, T., Martín-Vide, C., Pironkov, G. (szerk.) Statistical Language and Speech Processing. pp. 91–102. Springer International Publishing, Cham (2018)
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>

Smooth inverse frequency based text data selection for medical dictation

Domonkos Bálint^{1,2}, Péter Mihajlik^{1,3}

¹Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics, 1111, Budapest,
Műegyetem rkp. 3.

balintdomonkosjozsef@edu.bme.hu, mihajlik@tmit.bme.hu

²SpeechTex Kft., 1181, Madách Imre u. 47.

³THINKTech Nonprofit Kft., 2600, Vác, Váczy Pál u. 15.

Abstract. Under-resourced domain problem is significant in automatic speech recognition, especially in small languages such as Hungarian or in fields where data is often confidential such as finance and medicine. We introduce a method using word embedding and smooth inverse frequency (SIF) based distance measurement to filter public domain web corpora. The selection for (medical) domain matching documents can be scaled. The resulted text is used to train an augmented language model for a medical dictation system. We show that using the appropriately scaled selection leads to optimal performance of the ASR system over the base-lines where no data augmentation was applied or all the augmentation data was added.

Keywords: data selection, data acquisition, smooth inverse frequency, automatic speech recognition, sentence embedding

1 Introduction

In automatic speech recognition (ASR) - as well as in every machine learning field - we need data that fits well to the later area of usage. In some cases (eg. in healthcare, financial) such data are very difficult and costly to collect, and even if they are available, their amount is far below what is required.

In a typical ASR system there are two main models that require data to train them, the acoustic and the language model. The former is trainable with non-domain-specific (general) data, however the latter cannot produce output words that are not included in the training set, so domain-specific data is essential. Overall this phenomenon results in lower accuracy speech recognition systems, with narrow usability.

To tackle this problem, the usual approach is to add general data to the training set that is not related to the targeted topic. However, this method typically only slightly improves the accuracy of speech recognition, while also can multiply the size of the models and thus the resource requirements.

This paper explores whether general data can effectively be filtered using machine learning methods and domain-specific training data, and whether the filtered data can be used as training data to increase accuracy of an ASR system. The rest of the paper is organised as follows. In Section 2 we wrote about the different data acquisition techniques used during ASR development. Section 3 contains information about the data sets employed here. Section 4 continues with the settings of the experiments carried out in this paper, including the different word embedding models explored and the word embedding aggregation method used. The different settings were evaluated in an independent test set, the results are presented in Section 5. The paper finishes with conclusions in Section 6.

2 Related Works

We can distinguish between three main approaches, addressing this under-resourced domain problem, such as text generation and augmentation, text translation and data crawling. We will briefly introduce them in that order.

Text generation using recurrent neural network (RNN) architectures is a common application that can be used to address this problem (Barzilay and Lapata, 2005), (Koncel-Kedziorski et al., 2019). Creating text of similar nature to the available limited amount of domain specific data is one straightforward way to increase the size of the corpus. Transformer networks are also used for this task (Tarján et al., 2020), as they tend to provide state-of-the-art results in the field of natural language processing (NLP) and especially in text generation (Brown et al., 2020), (Devlin et al., 2018).

One different approach of data augmentation was shown in (Wei and Zou, 2019). Through simple operations like synonym replacement, random insertion, random swap, and random deletion this was able to improve the performance of the examined neural networks on five different NLP tasks.

In the case of end-to-end ASR (E2E) it has been shown that a back-translation (Sennrich et al., 2015), (Lample et al., 2017) style data augmentation could improve its performance (Hayashi et al., 2018). An E2E ASR system is typically an encoder-decoder neural network, and it is only trainable with voice, text pairs (Cho et al., 2014), (Sutskever et al., 2014). This method can utilize unpaired text corpora, as they are used to train only the decoder network.

Language models trained on a machine translated text have been found to be useful in a low-resourced setting (Jensson et al., 2008). Cross-lingual language model pretraining could also effectively be used for under-resourced languages (Lample and Conneau, 2019). Overall we found that medical data is confidential in every developed country, therefore it is similarly difficult to acquire clean data in e.g. English as in Hungarian.

Many publications refer to online text data acquisition as a possible solution (Sethy et al., 2006), (Remus and Biemann, 2016a). One approach is focused web-crawling (Chakrabarti et al., 2000). This term refers to the process of crawling the web in a guided way with focus on a specific topic. Without any labeled data, one proven method is language model and perplexity based crawling (Remus and

Biemann, 2016b). In (Vogel et al., 2020) they used different document similarity measuring methods to automatically collect in-domain texts from the web.

Our work differs from existing approaches as we used traditional crawling to acquire data, and filtered this data to get the final training set. In this way we could find the closest documents to our domain, therefore we did not have to set a hard limit to determine which documents were close enough to our reference. In addition, to the best of our knowledge this kind of document similarity based database creation is not yet published for Hungarian language.

3 Data Sources

The experiments presented in this paper were carried out using several datasets which will be explained in the next subsections.

WebBeteg: To evaluate the accuracy of our models, we needed a database that contained topic-specific data but also many other irrelevant data. For this, we selected the medical question-and-answer section of the WebBeteg Hungarian healthcare site¹.

The resulting database contains nearly 150 000 questions, of which 10% contain medical records (based on manual sampling and evaluation).

From this database, we manually selected 50 documents that contained medical records. We have considered this as the reference set. Our goal was to find similar entries in the rest of the database. This set contains 3256 tokens, its vocabulary size is 1751. As we used our reference set as a test set, we created audio files from the selected texts. The text audio is 36 minutes long.

Proprietary data: A large, but to our specific task only loosely related medical database was already available. This includes medical journals, medication descriptions, and a small amount of X-ray records. The database consists of 1 717 851 unique tokens, overall its size is 60 362 655 tokens.

4 Methods

4.1 Word Embeddings

In order to use words in machine learning models, they have to be represented with a numerical form. Over the years researchers have used many word representations like bag-of-words, one-hot encoded vectors etc. However the recent neural models like word2vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014) provide better representations to the words considering its context, too. Their main weakness is that every word has a unique word embedding regardless of the context it appears. As an example the word 'bank' in two sentences - "I

¹ <https://www.webbeteg.hu/orvos-valaszol>

am walking by the river bank” and “I deposited money to the bank” would have the same embeddings which can be confusing for machine learning models. The recent introduction of contextualised word representations solved this problem by providing vectors for words considering their context too. In this way the word ‘bank’ in the above sentences has two different embeddings. As a result, contextualised word embeddings perform better than standard word embeddings in many natural language processing tasks like question answering, textual entailment etc. (Devlin et al., 2018). The following words representation models were considered for the experiments.

FastText FastText is a library for learning word embeddings based on (Bojanowski et al., 2017). This provides a model, which is based on the skip-gram model, where each word is represented as a bag of character n-grams. A vector representation is associated with each character n-gram; words being represented as the sum of these representations. This representation has a very useful effect on a corpus with many rare words (eg. corpora written in an agglutinative language, or a corpus with a lot of misspelled words): similar words have similar representations, if they are in the same context. We used a pretrained model provided in the official fasttext website.² It is a 300 dimension embedding, and it was trained on the Hungarian Wikipedia.

ELMo ELMo introduced by (Peters et al., 2018) uses bidirectional language model (biLM) to learn both word (e.g., syntax and semantics) and linguistic context. After pretraining, an internal state of vectors can be transferred to downstream natural language processing tasks. We used a pre-trained Hungarian model provided in (Che et al., 2018), (Fares et al., 2017) which trained on a 20 million sample of WikiDump³ and Common Crawl⁴. Using the model we represented each word as a vector with a size of 4096 values.

4.2 Smooth Inverse Frequency

We acquired sentence embeddings using Smooth Inverse Frequency (SIF) proposed by (Arora et al., 2017) and then calculated the cosine similarity between those embeddings.

Semantically speaking, taking the average of the word embeddings in a sentence tends to give too much weight to words that are quite irrelevant. Smooth Inverse Frequency tries to solve this problem in two steps.

- Weighting: Smooth Inverse Frequency takes the weighted average of the word embeddings in the sentence:

$$v'_s = \frac{a}{|s|} \sum_{w \in s} \frac{a}{a + P(w)} \cdot v_w \quad (1)$$

² <https://fasttext.cc/docs/en/pretrained-vectors.html>

³ <https://dumps.wikimedia.org/>

⁴ <https://commoncrawl.org/>

where s is the input sentence, w is a word in s , v_w is the word embedding of w , $P(w)$ is the estimated frequency of w and a is a parameter that is typically set to 0.001.

- Common component removal: We assume that there is n sentences in the corpus, in the next step SIF creates a matrix from all the previously calculated sentence embeddings:

$$X = [v'_{s1}|v'_{s2}|\dots|v'_{sn}] \quad (2)$$

Then the algorithm computes the principal component of X . It then subtracts their projections on first principal component from these sentence embeddings:

$$v_s = v'_s - uu^T v'_s \quad (3)$$

where u is the principal component of X . This should remove variation related to frequency and syntax that is less relevant semantically. v_{s1} is the final sentence embedding output for sentence $s1$.

As a result, Smooth Inverse Frequency downgrades unimportant words such as *but*, *just*, etc., and keeps the information that contributes most to the semantics of the sentence. After acquiring the sentence embeddings for a pair of sentences, the cosine similarity between those two vectors were taken to represent the similarity between them.

In (Ranasinghe et al., 2019) they showed that SIF with the previously presented word embeddings can perform in the same level as the much more complex transformer networks on the task of document similarity, therefore we chose these methods in our paper.

4.3 Language modeling

N-gram models Back-off, n-gram language models (BNLMs) are still commonly used in online, single-pass speech transcription systems due to their lower source demand and high compatibility with Weighted Finite-State Transducer (WFST) decoders. Hence we applied BNLMs as our language models in our experiments. All BNLMs are trained with modified Kneser-Ney discounting (Chen and Goodman, 1996) applying the implementation of SRI language modeling toolkit (Stolcke, 2002). We carried out a preliminary experiment on the development set and found 3-gram the optimal LM order for word BNLMs.

5 Results

This section describes the evaluation results of WebBeteg data for all methods we mentioned above. All experiments were evaluated using WER (word error rate), and LER (letter error rate) of the ASR system. We also calculated the perplexity and the out-of-vocabulary (OOV) word count for every language model. The latter gives us a more precise image of the goodness of our text filtering method, as it does not include the noise from the acoustic model. However during our

evaluation we considered WER and LER more important metrics, as our final goal was to improve the ASR system and these gave us information about the accuracy of the whole ASR system, not just the language model.

A new language model was trained for every setup, but the acoustic model and the decoding remained the same. The latter two are described thoroughly in (Varga et al., 2015). The following subsections will discuss the results in detail in each case.

5.1 FastText and SIF

We assigned a vector to the texts in the WebBeteg database and to the reference text using the SIF algorithm (Arora et al., 2017) and FastText word embedding (Bojanowski et al., 2017). We then calculated the cosine similarity between the text vectors and the reference vector, and established an order based on these distances. Then, we selected the first n pieces from the database based on the order. The names of the models were created from this n number: for example, in the FastText f100 model, we selected the first 100 items from the WebBeteg database. Table 1 shows the size of the resulting text sets. We created a language model from it, and interpolated with the language model from the proprietary database (see Section 3). The interpolation weights of the two models were 0.5, 0.5, based on a one variable optimization process.

dataset	Token count	Unique token count
Proprietary data	60 362 655	1 717 851
WebBeteg	13 068 749	945 280
WebBeteg f100	20 310	6 802
WebBeteg f1 000	158 851	33 171
WebBeteg f10 000	1 228 734	181 089
WebBeteg f100 000	9 138 986	700 208

Table 1. The token count, and the unique token count of the different databases. (e.g. WebBeteg f100 is the first 100 best text from WebBeteg database - see section 5.1)

Table 2 shows the results for models using the data sorted by FastText and SIF methods as we described above.

5.2 ELMo and SIF

We followed the same algorithm as described in Section 5.1, with the difference that ELMo embedding (Peters et al., 2018) was used instead of FastText embedding.

The naming conventions are the same as before: for example, in ELMo f100 we selected the first 100 elements of the WebBeteg database to train the model. The results are shown in the Table 2.

model name	size	WER (%)	LER (%)	perplexity	OOV (%)
baseline1		27.14	8.36	580	3.34
baseline2		23.34	6.92	406	3.14
FastText	f100	24.03	7.08	398	3.34
	f1 000	22.86	6.57	295	3.34
	f10 000	23.14	6.98	278	3.21
	f100 000	23.21	6.95	379	3.14
ELMo	f100	24.28	7.15	402	3.34
	f1 000	23.28	6.78	279	3.31
	f10 000	22.97	6.79	268	3.24
	f100 000	23.34	6.92	392	3.14

Table 2. The WER, LER, perplexity and OOV results of the different models described in Section 5. Baseline1 and baseline2 are when we add nothing or everything from WebBeteg respectively.

6 Conclusion

In this paper, we filtered an online database to create a smaller in-domain set. We examined whether the WER and LER values of a speech recognizer can be improved by interpolating the language model created from this textset with another language model trained on an orders of magnitude larger database, thereby adapting the resulting language model to the task.

The results (Table 2) show that the recognition metrics of the models after one point begin to decrease with the additional data. This is advantageous as we can improve the model without a huge increase in complexity or size. We also established that a more complex word embedding like ELMo (Peters et al., 2018) can’t improve the aforementioned filtering. The reason why there is no significant difference between results of the different embedding techniques applied requires further investigations. Overall we found it useful to perform text similarity based filtering for noisy databases used in speech recognition training.

Bibliography

- Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: ICLR (2017)
- Barzilay, R., Lapata, M.: Collective content selection for concept-to-text generation. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 331–338. Association for Computational Linguistics, Vancouver, British Columbia, Canada (Oct 2005), <https://www.aclweb.org/anthology/H05-1042>
- Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5, 135–146 (2017)
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)
- Chakrabarti, S., Berg, M., Dom, B.: Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks* 31, 1623–1640 (04 2000)
- Che, W., Liu, Y., Wang, Y., Zheng, B., Liu, T.: Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 55–64. Association for Computational Linguistics, Brussels, Belgium (October 2018), <http://www.aclweb.org/anthology/K18-2005>
- Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting on Association for Computational Linguistics. p. 310–318. ACL ’96, Association for Computational Linguistics, USA (1996), <https://doi.org/10.3115/981863.981904>
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation (2014)
- Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018), <http://arxiv.org/abs/1810.04805>
- Fares, M., Kutuzov, A., Oepen, S., Velldal, E.: Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In: Proceedings of the 21st Nordic Conference on Computational Linguistics. pp. 271–276. Association for Computational Linguistics, Gothenburg, Sweden (May 2017), <http://www.aclweb.org/anthology/W17-0237>
- Hayashi, T., Watanabe, S., Zhang, Y., Toda, T., Hori, T., Astudillo, R., Takeda, K.: Back-translation-style data augmentation for end-to-end asr (2018)
- Jensson, A., Iwano, K., Furui, S.: Language model adaptation using machine-translated text for resource-deficient languages. *EURASIP J. Audio Speech Music Process.* 2008(1) (Dec 2008)

- Koncel-Kedzierski, R., Bekal, D., Luan, Y., Lapata, M., Hajishirzi, H.: Text generation from knowledge graphs with graph transformers (2019)
- Lample, G., Conneau, A.: Cross-lingual language model pretraining (2019)
- Lample, G., Conneau, A., Denoyer, L., Ranzato, M.: Unsupervised machine translation using monolingual corpora only (2017)
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. CoRR abs/1310.4546 (2013), <http://arxiv.org/abs/1310.4546>
- Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014), <https://www.aclweb.org/anthology/D14-1162>
- Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. CoRR abs/1802.05365 (2018), <http://arxiv.org/abs/1802.05365>
- Ranasinghe, T., Orasan, C., Mitkov, R.: Enhancing unsupervised sentence similarity methods with deep contextualised word representations. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). pp. 994–1003. INCOMA Ltd., Varna, Bulgaria (Sep 2019), <https://www.aclweb.org/anthology/R19-1115>
- Remus, S., Biemann, C.: Domain-specific corpus expansion with focused webcrawling. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). pp. 3607–3611. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016a), <https://www.aclweb.org/anthology/L16-1572>
- Remus, S., Biemann, C.: Domain-specific corpus expansion with focused webcrawling. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16). pp. 3607–3611. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016b), <https://www.aclweb.org/anthology/L16-1572>
- Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data (2015)
- Sethy, A., Georgiou, P.G., Narayanan, S.: Text data acquisition for domain-specific language models. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. pp. 382–389. Association for Computational Linguistics, Sydney, Australia (Jul 2006), <https://www.aclweb.org/anthology/W06-1645>
- Stolcke, A.: Srilmm - an extensible language modeling toolkit. In: INTERSPEECH (2002)
- Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks (2014)
- Tarján, B., Szaszák, G., Fegyó, T., Mihajlik, P.: Deep transformer based data augmentation with subword units for morphologically rich online asr (2020)
- Varga, Á., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T., Bordás, C., Mihajlik, P.: Automatic close captioning for live hungarian television broadcast speech:

- A fast and resource-efficient approach. In: Ronzhin, A., Potapova, R., Fakotakis, N. (eds.) *Speech and Computer*. pp. 105–112. Springer International Publishing, Cham (2015)
- Vogel, I., Choi, J.E., Meghana, M.: Similarity detection pipeline for crawling a topic related fake news corpus (2020)
- Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks (2019)

Automatikus hibajavítás statikus szövegeken

Gulás Máté^{1,2}, Yang Zijian Győző¹, Dömötör Andrea^{1,2}, Laki László János¹

¹MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
1083 Budapest, Práter u. 50/a.

²Pázmány Péter Katolikus Egyetem Bölcsész- és Társadalomtudományi Kar
2087 Piliscsaba, Egyetem u. 1.
gulas.mate@hallgato.ppke.hu,
{yang.zijian.gyozo, domotor.andrea, laki.laszlo}@itk.ppke.hu

Kivonat Kutatásunk célja egy olyan neurális hálózat alapú automatikus hibajavító eszköz létrehozása, amely képes a korpuszok sztenderdizálására. A különböző nyelvtechnológiai feladatok modelljeinek betanításához fontos, hogy a tanítókorpuszok minél kevesebb zajt illetve hibát tartalmazzanak, hiszen a gyenge minőségű tanítókorpuszok rendszerint rosszabb eredményekhez vezethetnek. Az interneten elérhető szövegek nagy része informális, nem ellenőrzött forrásból (pl. közösségi média, fórumok) származik. Tanulmányunkban a közösségi médiában gyakran előforduló gyakori hibákra fókuszálunk. Célunk feltárni és elemezni a hibatípusokat, majd az előfordulásuk alapján statisztikát készíteni. A kiszámolt hiba-előfordulások arányát felhasználjuk egy hibajavító modell tanítására. Kutatásunkban egy transzformer modellen alapuló neurális gépi fordító rendszert használtunk fel a hibajavító modell tanítására. Eredményeink azt mutatják, hogy a neurális gépi fordítás módszere alkalmas a feladatra, azonban több olyan hibatípus is létezik, amelyek további kutatást igényelnek.

Kulcsszavak: hibajavítás, hiba-előfordulás, korpusztisztítás, neurális gépi fordítás

1. Bevezetés

Napjaink számítógépes nyelvészeti megoldásainak nagy részéhez elengedhetetlenek a nagyméretű korpuszok. A különböző nyelvtechnológiai feladatok modelljeinek betanításához fontos, hogy a tanítókorpuszok minél kevesebb zajt illetve hibát tartalmazzanak, hiszen a gyenge minőségű tanítókorpuszok rendszerint rosszabb eredményekhez vezetnek. Eredendően hibátlan korpuszt találni szinte lehetetlen, hiszen az interneten elérhető szövegek nagy része informális, nem ellenőrzött forrásból (pl. közösségi média, fórumok) származik, de még az ellenőrzött szövegekben is lehet hibákat találni.

Ennek a problémának a megoldására egy neurális gépi fordítás (NMT) alapú hibajavítót javasunk. Ennek segítségével egyszerűen tudunk szövegtörzseket javítani. Hibajavítónk magában foglalja az előforduló hibák széles skálájának javítását.

2. Kapcsolódó irodalom

Magyar nyelvre Dömötör és Yang (2018) végeztek kutatást a nem sztenderd hibák előfordulására korpuszokban. Első sorban beszélt nyelvi és a személyes alkörpuszokból indultak ki, hiszen ott lehet nagyobb mennyiségben hibás szöveget találni. Kimutatták, hogy a nem sztenderd hibák közel 30%-át kiugróan az írásjelek és nagybetűk elhagyása teszi ki. 15% körül van az elütések, helyesírási és nyelvi hibák, 10% körül a nehezen értelmezhető beszélnyelvi szöveg, kicsivel kevesebb, de ugyancsak 10% körül az ékezetek hiánya, 5% körül a szegmentálási hibák és végül kevesebb mint 2% körül az idegen szavak. Kutatásunkban ezeket a hibákat vizsgáltuk meg még részletesebben, kifejezetten közösségi médiákból származó kommentek és szövegek között.

A hibák visszaállításának gépi fordítással történő első megközelítése Brockett és mtsai (2006) nevéhez fűződik. Tanulmányukban statisztikai gépi fordítást (SMT) használtak azzal a céllal, hogy az angolt második nyelvként tanulók főnevekkel kapcsolatos hibáit javítsák. Az interneten nagy számban előforduló főnévi hibák 61,81%-át sikerült javítaniuk módszerükkel.

Az ő megközelítésükre alapozva többek között Felice és mtsai (2014) egy szabály és SMT alapú hibajavító rendszert alkottak, míg Susanto és mtsai (2014) egy klasszifikációs és SMT alapú rendszer kombinációját mutatták be.

Ahogy a gépi fordítás területén a mondataalapú NMT egyre jobb eredményeket produkált a kifejezés-alapú SMT-vel szemben, a hibajavítás terén is egyre több NMT alapú megközelítés született, kezdetben azonban a kifejezés-alapú SMT rendszerek felülmúlták az NMT rendszereket. Az első jelentős NMT alapú hibajavító rendszert Junczys-Dowmunt és mtsai (2018b) mutatták be. Kutatásukban az alacsony-erőforrású NMT-k számos módszerét ültették át NMT alapú hibajavító rendszerükbe és további modell-független módszereket vezettek be. Hibajavító rendszerük több mint 10%-kal jobb eredményt produkált, mint az addigi state-of-the art NMT alapú hibajavító, és 2%-kal jobb eredményt ért el a CoNLL-2014 benchmark szerint, mint az addigi nem-neurális state-of-the-art rendszer.

Egy további úttörő NMT alapú hibajavító rendszert mutatott be Chollam-patt és Tou Ng. (2018), akik egy többrétegű, konvolúciós encode-decoder neurális hálót használtak. A hálót olyan beágyazásokkal inicializálták, melyek felhasználják a karakter n-gram információt.

A közelmúltban nagyon népszerű lett az NMT alapú hibajavítás, több kutatás is született a témában. Ezek közül érdemes kiemelni Zhao és mtsai (2019) munkáját, akik egy másolási mechanizmust vezettek be: a változatlan és a szótáron kívüli szavakat közvetlenül átmásolták a forrásmondatból a célmondatba.

Jelenleg nincs tudomásunk arról, hogy magyar szerzők foglalkoztak volna NMT alapú hibajavítással, vagy külföldi szerzők magyar nyelvű NMT hibajavító rendszerrel.

3. Hibatípusok informális szövegekben

A hibajavító első lépéseként az volt a célunk, hogy felmérjük, milyen típusú hibák fordulnak elő informális szövegekben. Ehhez a TrendMiner projekt korpuszát¹ használtuk (Miháltz és mtsai, 2015). A korpusz 1,9 millió magyar nyelvű, politikai témájú Facebook hozzászólást tartalmaz morfológiai elemzéssel együtt. A korpusz kiválóan alkalmas a hibák keresésére, hiszen egyrészt a közösségi média kommentek nem szerkesztett, nem ellenőrzött mondatokból állnak, másrészt a morfológiai elemzés Out of Vocabulary (OOV) címkéje segíti a hibás szövegek keresését. A szövegek jellemző hibáinak feltérképezésére tehát azokat a kommenteket használtuk fel, amelyekben szerepel legalább egy szó OOV címkével. Az így automatikusan kiválasztott kommentekből 350-et vizsgáltunk meg részletesen. Ezekben számos olyan hibát is találtunk, amelyeket a morfológiai elemzés nem jelölt ismeretlen szónak.

A talált hibatípusokat három fő kategóriába osztottuk be aszerint, hogy milyen nyelvi elemzési szint szükséges a generálásukhoz. Az így kapott három kategória és a hibák típusai az 1–2. táblázatokban láthatók.

1. Felszíni alakból generálható

Ékezetek hiánya
 Kisbetű/nagybetű tévesztés
 Írásjelek hiánya, hibái
 Magánhangzó hosszúság/rövidség
 Mássalhangzó hosszúság/rövidség
 Plusz vagy hiányzó karakter
 Dátumok, számjegyes kifejezések hibái
 j-ly tévesztés
 Informális rövidítések (*h*, *vmi*, stb.)
 Egyéb gyakori speciális hibák: pl. *-ban/-ba*, *sem/se*, *lesz/lessz*
 Elgépelések

1. táblázat. A TrendMiner korpusz gyakori hibatípusai (felszíni alakból generálható hibák)

Jelen munkánkban nem foglalkozunk a morfológiai és a szintaktikai elemzés után generálható hibákkal. Ennek oka többek között az, hogy egyfelől az egyszerű/felszíni alakból generálható hibák alkotják a talált hibatípusok legnagyobb részét, másrészt ezeket a típusú hibákat viszonylag gyorsan és könnyen elő lehet állítani.

¹ <http://corpus.nyttud.hu/trendminer/>

Elemzés szükséges	
2. Morfológia	3. Szintaxis
Összetett szavak különírása	Hibás egyeztetés
Tagadószó egybeírása a következő szóval	Ragozás hibái vagy hiánya
Kopula egybeírása az előző szóval	
Jelzős szerkezetek egybe-/különírása	
Módosító szerkezetek egybe-/különírása	
Véletlenszerű egybeírás (szóközhány)	
Magánhangzó-harmónia megsértése	
Suksükölés	
Igekötők egybe-/különírása	

2. táblázat. A TrendMiner korpusz gyakori hibatípusai (elemzéssel generálható hibák)

4. Módszer

4.1. A tanítókorpusz előállítása

Második lépésként előállítottunk egy olyan párhuzamos korpuszt, amely mindkét oldalán ugyanazokat a mondatokat tartalmazza, annyi különbséggel, hogy az egyik oldalán ugyanannak a mondatnak a helyes, míg a másikon a hibás változata található. A tanítókorpusz létrehozásához az online elérhető Open Subtitles² nevű angol-magyar párhuzamos korpuszának magyar oldali szövegét használtuk. A korpusz TV és mozi filmekre létrehozott feliratokból áll. Ennek megfelelően főleg rövidebb, informális mondatokat tartalmaz. A tanítókorpusz 1 millió szegmensből, a validációs korpusz 5000 mondatból és a tesztkorpusz 3000 mondatból áll. A korpusz az egyik legnagyobb szabadon hozzáférhető párhuzamos tanítóanyagának számít, ellenben mérete elmarad az egynyelvű tanítóanyagokétól. Választásunk azért esett erre az adathalmazra, mert több párhuzamos kutatásunk során is használjuk, és néhány korpusztisztító lépést már előzetesen eszközöltünk rajta. Kivettük azokat a mondatokat, amelyek speciális karaktereket (pl. kínai, japán, cirill stb.) tartalmaztak, valamint a teszt halmaz mondatait kézzel kijavítottuk. Mérete elégséges a neurális hálózatok helyes betanítására, valamint a tanítási idő is viszonylag kezelhető marad (1-2 nap).

A tanítókorpuszunk forrásnyelvi oldalán található (feltételezett) helyes mondatokat automatikusan „rontottuk el” az 1. táblázatban található, felszíni alakból generálható hibákkal. Az így kapott mondatok alkotják a párhuzamos korpuszunk hibás oldalát. Kutatásunkban nem használtunk morfológiai és szintaktikai elemzőt, ezért az automatikus elrontó szkriptünk nem állít elő morfológiai és szintaktikai hibákat. Az automatikus hibageneráló szkriptünk a 3. táblázat arányai alapján ront el szavakat, az egyes hibatípusok ugyanis ebben az arányban fordultak elő a vizsgált korpuszban.

Miután lefuttattuk a szkriptünket az 1 millió mondat (6.181.241 token) szavaira, az alábbi hibaarányt kaptuk a mesterségesen előállított korpuszban:

² <http://opus.nlpl.eu/OpenSubtitles-v2018.php>

- szavak 6,55%-a hibás (egy szó több hibát is tartalmazhat)
- mondatok 32,04%-a hibás

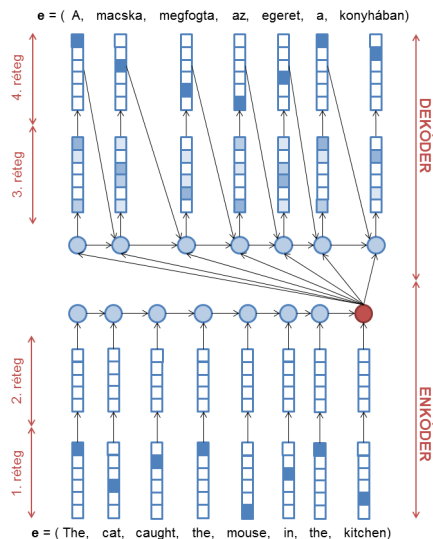
Hiba típusa	arány
Ékezetek hiánya	2,758%
Mondatkezdő kisbetű	9,43%
Mondatvégi írásjel hiánya	6,82%
Hosszú magánhangzó helyett rövid	6,84%
Rövid magánhangzó helyett hosszú	0,245%
Hosszú mássalhangzó helyett rövid	1,87%
Magánhangzó nyúlás l előtt (pl. <i>hátul</i>)	1%
Vonatkozó névmások előtti vessző hiánya	22%
„hogy” előtti vessző hiánya	24%
„ly” betű helyett „j”	4,34%
„lesz” szó helyett „lessz”	9%
Extra pont	7,14%
Extra vessző	10,85%
Informális rövidítések	3%
Extra karakterek beszúrása	0,12%
Hiányzó karakter	10,85%
-ban, -ben helyett -ba, -be	1,42%

3. táblázat. Hibák aránya

4.2. A neurális gépi fordítórendszer

A 2010-es évek első felére a statisztikai gépi fordítórendszerek elérték teljesítőképességük határát. Az alapjait képező módszert és a létrehozott keretrendszereket a kutatók nagyon sok befektetett munka ellenére lényegében nem sikerült tovább javítani. Az áttörést (Bahdanau és mtsai, 2015) rendszere hozta el, ami egy figyelmi (attention) modellel támogatott enkóder-dekóder architektúrájú NMT rendszer volt. A modell lényege, hogy kettéválasztja a fordítás folyamatát két elkülöníthető részre. A kódolás során lényegében egy RNN-alapú seq2seq modellel hoz létre, tehát a szóbeágyazási modellhez hasonlóan a fordítandó modellekből egy n -dimenziós vektort készít. Az 1. ábrán ez a vektor felel meg az ábra közepén látható piros/sötét node-nak. A második fázis a dekódolás, ahol a mondatvektorból generálja ki a célnyelvi mondatot egy RNN réteg segítségével.

Innentől számítva az NMT rendszerek átvették a vezető szerepet az SMT-től. 2017-ben a Google cég munkatársai (Vaswani és mtsai, 2017) publikálták és szabadon hozzáférhetővé tették az úgynevezett multi-attention réteggel támogatott NMT rendszerüket. Ezt a szakirodalomban transzformer-alapú architektúrának nevezik. A módszer lényege, hogy az eddigi egy helyett több figyelmi réteget helyeztek el a rendszerben, ami segítségével nagymértékben nőtt a többértelmű szavak fordításának minősége.



1. ábra: Enkóder-dekóder architektúra vázlatos rajza

Munkánk során a Marian NMT (Junczys-Dowmunt és mtsai, 2018a) nevű keretrendszert használtuk, ami egy C++ nyelven íródott szabadon hozzáférhető programcsomag. Könnyen telepíthető, jól dokumentált, memória- és erőforrás-optimalis implementációjának köszönhetően³ az akadémiai felhasználók és fejlesztők által leggyakrabban használt eszköz (Barrault és mtsai, 2019).

4.3. A Sentence Piece tokenizáló

Az NMT rendszerek működése GPU processzorokon történik, melyek egyik szűk keresztmetszete a bennük található memória mérete. Ez határozza meg a létrehozható NMT rendszer szótárának a méretét. Egy szóalapú rendszer esetében az általánosságban 100K különálló szóban korlátozzák le a rendszert, így a további szavakat ismeretlenként kezeli.

(Sennrich és mtsai, 2015) ezt a problémát úgy oldották meg, hogy a szavak helyett úgynevezett subword (szótöredék) szintre csökkentették a legkisebb fordítási egységet. A BPE (Byte Pair Encoding) egy adattömörítő eljárás, ahol a leggyakoribb bájt párokat egy olyan bájjal helyettesítjük, amely nem szerepel magában az adatban. Az eljárás a korpuszon először egy karakteralapú szótárat hoz létre, ahol minden szót karakterek sorozataként ábrázol. Ezután gyakoriság alapján a gyakori karaktersorozatokat önálló tokenekként kezeli. Ezzel az adat tömörítése mellett az ismeretlen szavak kezelését is megoldja, hiszen a

³ <https://marian-nmt.github.io/>

részsavakból előállítható egy olyan összetétel, amely nem szerepelt eredetileg a korpuszban.

Ezt a módszert fejlesztették tovább (Kudo és Richardson, 2018). Az általuk létrehozott Sentence Piece nevű eszköz egy felügyelet nélküli szöveg tokenizáló és detokenizáló, melyet elsősorban a neurálishálózat-alapú gépi tanulási feladatokhoz fejlesztettek ki. Implementálva van benne a BPE metrika, ami egy unigram nyelvmodellel (Kudo, 2018) van súlyozva. Használatával elhagyhatók a költséges nyelvspecifikus előfeldolgozási lépések, mint például a tokenizálás vagy a kisbetűsítés. A módszer lényege, hogy a természetes szöveget úgy alakítja át, hogy abban a különböző „szavak” száma korlátos legyen, valamint az így létrejött tanítóanyagban nem lesznek ismeretlen szavak. Ennek köszönhetően a neurális hálózatok paraméterszáma nagymértékben csökkenthető.

- (1) Sima szöveg: Petőfi Sándor egy nagyszerű költő.
SPM szöveg: P ető fi □ S ándor □egy □nagyszerű □költő .

A fenti példában látható az SPM (Sentence Piece modell) kimenete. A sima szöveg szavait gyakran előforduló karakter sorozatokra tördeli szét. Érdekes megfigyelni, hogy az eredeti mondat szóközeit is a szavakhoz csatolja és mint önálló karaktert (□) kezeli.

4.4. NMT tanítása

Az NMT tanításához a Marian neurális gépi fordítórendszert használtuk. A rendszer fontos jellemzője a SPM technológia. A rendszerünk tanításához az alábbi hyper-paramétereket használtuk:

- Sentence Piece: szótárméret: 16000; egy szótár a forrás- és egy a célnyelvi korpusznak; karakter lefedettség a teszt korpuszon: 100%
- Transzformer modell: enkóder és dekóder rétegeinek száma: 6; transformer-dropout: 0,1;
- learning rate: 0,0003; lr-warmup: 16000; lr-decay-inv-sqrt: 16000;
- optimizer-params: 0,9 0,98 1e-09; beam-size: 6; normalize: 0,6
- label-smoothing: 0,1; exponential-smoothing

A gépi fordítás forrásnyelve a hibageneráló szkriptünk által „elrontott” tanítókorpusz, míg a célnyelv az eredeti nem hibás tanítókorpusz.

A tanítás körülbelül 7 óra alatt végzett.

5. Eredmények

A 4. táblázat eredményein megfigyelhető, hogy a fedés szinte minden esetben alulmarad a pontossághoz képest, azaz a rendszer inkább "óvatos", mint alapos. Ezt az is okozhatja, hogy a tanítóanyagban a mondatpárok hibátlanak tekintett oldala valójában nem volt mindig hibátlan. A legjobb eredményt a tipikusan az internetes fórumokra jellemző hibákkal értük el (ly-j, lesz-lessz, rövidítések).

	Pontosság	Fedés	F-mérték	Esetek száma (res - gold)
Nagybetű	92,35%	90,41%	91,37%	327 - 334
Ékezetes szavak	88,44%	81,55%	84,85%	545 - 591
Mondatvégi írásjelek	82,42%	78,18%	80,25%	387 - 408
Mondatvégi pont	84,75%	86,06%	85,40%	328 - 323
Mondatvégi kérdőjel	75,67%	60,86%	67,46%	37 - 46
Mondatvégi felkiáltójel	61,11%	34,37%	44,00%	18 - 32
Vessző	93,47%	90,76%	92,10%	368 - 379
Rövid MGH	94,07%	79,37%	86,10%	135 - 160
Hosszú MSH	88%	68,75%	77,19%	50 - 64
Hosszú MGH	89,31%	70,05%	78,52%	131 - 167
Hosszú MGH l előtt	84,61%	70,96%	77,19%	26 - 31
ly-j	100%	92,85%	96,29%	13 - 14
lesz-lessz	100%	100%	100%	18 - 18
ban, ben	100%	60%	74,99%	3 - 5
rövidítések	100%	100%	100%	16 - 16

4. táblázat. Hibatípusok relatív eredményei

Ezek valószínűleg egyébként nem fordulnak elő az OpenSubtitles korpuszban, így valóban csak a hibás szöveg-generáló szkripttel kerülhettek a tanítóanyagba. A mondatvégi írásjelek pótlásában a felkiáltójel eltalálása bizonyult a legnehezebb feladatnak, ami nem meglepő, hiszen sok esetben az ember számára is nehéz eldönteni, hogy egy mondatot felkiáltónak szánt-e a szerzője. Ugyanez elmondható a kérdőjellel kapcsolatban is: az eldöntendő kérdések és a kijelentő mondatok között az írásjelen kívül nincs különbség. A vesszőhibákkal kapcsolatban meg kell említeni, hogy a rendszer csak a *hogy* és a vonatkozó névmások előtti vessző pótlására lett tanítva, ezeket viszonylag jó eredménnyel tudta teljesíteni. A vesszőhibák egyébként, amint a 3. táblázatban látható, az összes hiba csaknem felét teszik ki.

6. Összegzés

Kutatásunk célja egy olyan NMT alapú automatikus hibajavító eszköz létrehozása volt, amely képes a korpuszok sztenderdizálására. Ebben a tanulmányban a közösségi médiában gyakran előforduló gyakori hibákra fókuszáltunk. A korpuszméréseink szerint a legjellemzőbb hibatípus a vesszőhiány, rendszerünk ezt 90% körüli pontossággal és fedéssel tudja javítani. Hibajavítónk ezen kívül eredményes volt néhány tipikus, informális szövegekre jellemző hiba javításában, a kevésbé specifikus hibák esetén azonban nem volt annyira sikeres. Ennek oka lehet a tanítókorpusz "helyes" oldalának esetlegesen rossz minősége, illetve az is, hogy a generált hibák nem minden esetben eléggé „élethűek”.

A további feladataink között szerepel ezért a tanítókorpusz minőségének javítása, illetve a hibajavító kiterjesztése további, már elemzést igénylő hibatípusokra is. Továbbá a neurális gépi fordítás mellett szeretnénk kipróbálni az új

kontextuális szóbeágyazás-alapú modellek, mint a BERT jellegű modellek alkalmazását is.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1-NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

- Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (szerk.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), <http://arxiv.org/abs/1409.0473>
- Barrault, L., Bojar, O., Costa-jussà, M.R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Mázler, M., Pal, S., Post, M., Zampieri, M.: Findings of the 2019 conference on machine translation (wmt19). In: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). pp. 1–61. Association for Computational Linguistics, Florence, Italy (August 2019), <http://www.aclweb.org/anthology/W19-5301>
- Brockett, C., Dolan, W.B., Gamon, M.: Correcting esl errors using phrasal smt techniques. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. pp. 249–256. Association for Computational Linguistics, Stroudsburg, USA (2006)
- Chollampatt, S., Tou Ng, H.: A multilayer convolutional encoder-decoder neural network for grammatical error correction. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (2018)
- Dömötör, A., Yang, Z.G.: Így írtok ti: nem sztenderd szövegek hibatípusainak detektálása gépi tanulós módszerrel. XIV. Magyar Számítógépes Nyelvészeti Konferencia pp. 305–316 (2018)
- Felice, M., Yuan, Z., Andersen, O., Yannakoudakis, H., Kochmar, E.: Grammatical error correction using hybrid systems and type filtering. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task. pp. 15–24. Association for Computational Linguistics, Baltimore, Maryland (2014), <http://www.aclweb.org/anthology/W14-1702>
- Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Hermann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., Birch, A.: Marian: Fast neural machine translation in C++. In: Proceedings of ACL 2018, System Demonstrations. pp. 116–121. Association for Computational Linguistics, Melbourne, Australia (Jul 2018a)

- Junczys-Dowmunt, M., Grundkiewicz, R., Guha, S., Heafled, K.: Approaching neural grammatical error correction as a low-resource machine translation task. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pp. 595–606. Association for Computational Linguistics, New Orleans, Louisiana (2018b)
- Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 66–75. Association for Computational Linguistics, Melbourne, Australia (Jul 2018)
- Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018), <https://www.aclweb.org/anthology/D18-2012>
- Miháltz, M., Váradi, T., Csertő, I., Fülöp, É., Pólya, T., Kővágó, P.: Beyond sentiment: Social psychological analysis of political facebook comments in hungary. In: *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2015)*. pp. 127–133. Association for Computational Linguistics, Lisboa, Portugal (2015)
- Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. CoRR abs/1508.07909 (2015), <http://arxiv.org/abs/1508.07909>
- Susanto, R.H., Phandi, P., Tou Ng, H.: System combination for grammatical error correction. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 951–962. Association for Computational Linguistics (2014), <https://doi.org/10.3115/v1/D14-1102>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (szerk.) *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc. (2017), <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- Zhao, W., Wang, L., Shen, K., Jia, R., Liu, J.: Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (2019)

Szó, beszéd – avagy hogyan kommunikálunk egymásról

Üveges István^{1,2}, Szabó Martina Katalin^{2,3}, Vincze Veronika⁴

¹Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola
6722 Szeged, Egyetem utca 2.

²Szegedi Tudományegyetem, Informatikai Intézet
6720 Szeged, Árpád tér 2.

³Társadalomtudományi Kutatóközpont, CSS-RECENS
1097 Budapest, Tóth Kálmán utca 4.

⁴MTA-SZTE Mesterséges Intelligencia Kutatócsoport
6720 Szeged, Tisza Lajos körút 103.
uvegesistvan898@gmail.com
{martina,vincze}@inf.u-szeged.hu
Szabo.Martina@tk.hu

Kivonat A dolgozatban egy nagy méretű, magyar beszélt nyelvi adatbázis (HuTongue) segítségével azt vizsgáljuk, hogyan kommunikálnak egymásról a diskurzusok résztvevői, és megszólalásaikhoz milyen szentimentek és emóciók kapcsolódnak. A vizsgálathoz használt korpusz beszélt nyelvi szövegeket tartalmaz, amely manuálisan leiratozott és annotált. A korpusz létrehozásának elsődleges célja az volt, hogy a kutatócsoport elsősorban szociológiai kutatási kérdéseikhez megfelelő vizsgálati anyagot biztosítson. A csoport arra a kérdésre keresi a választ, hogy zárt közösségekben belül az informális kommunikáció és a megbecsültség hogyan befolyásolja a közösség működését és normarendszerét. A korpusz annotációja lehetővé teszi a csoporttagok egymásra vonatkozó közléseinek a vizsgálatát más szemantikai és pragmatikai sajátságokkal összefüggésben. A jelen dolgozatban kifejezetten ezekre az interakciókra fókuszálunk, és azokat a korpusz szentiment- és emóciótartalmaival összefüggésben elemezzük.

Kulcsszavak: magyar, diskurzus, korpuszelemzés, nyelvi jellemzők, kommunikáció, hálózat, hálózatelemzés, csoportkohézió, pletyka, szentiment- és emócióelemzés

1. Bevezetés

A dolgozatban a HuTongue korpusz alapján olyan magyar diskurzusok szemantikai és pragmatikai sajátságait tárgyaljuk, amelyben a résztvevők közlései valamely jelen nem levő személyre vagy személyekre irányulnak.

A HuTongue korpusz szövegei hétköznapi szituációkban, külső ingerektől elszigetelt környezetben keletkeztek (Galántai és mtsai, 2018; Gulyás és mtsai,

2018). A korpusz létrehozásának elsődleges célja az volt, hogy megfelelő kutatási anyagot nyújtson a pletyka elemzéséhez az emberi interakcióban, ugyanakkor a benne foglalt anyagok keletkezési körülményei, mennyisége és feldolgozásának módja miatt az adatbázis számos társadalomtudományi és nyelvészeti vizsgálat elvégzésére lehetőséget adhat.

Bár a pletyka fogalmának a szakirodalom alapján több, különféle megközelítése is ismert (Emler, 1994; Grosser és mtsai, 2012), a HuTongue korpusz annotálásához a résztvevő szakértők egy tág definíciót határoztak meg: az annotációs fázis során minden olyan megnyilatkozás ebbe a kategóriába kellett, hogy tartozzon, amelynek tárgya egy vagy több jelen nem levő személy volt. Ennek a meghatározásnak köszönhetően egy olyan korpusz jött létre, amely lehetővé teszi az emberi kommunikáció e szegmensének a széleskörű, valamely definíció által nem korlátozott vizsgálatát.

Az elmúlt évtizedekben több beszélt nyelvi korpusz keletkezett több nyelven (Crowdy, 1993; Hemphill és mtsai, 1990; Maekawa és mtsai, 2000; Oostdijk, 2000; Van Bael és mtsai, 2007). Neuberger és mtsai (2014) alapján, közöttük találunk agglutináló nyelveket reprezentáló korpuszokat is, például török (Mengusoglu és Deroo, 2001) és finn (Seppänen és mtsai, 2003). Azonban a nyilvánvaló változatosság ellenére is csak néhány olyan korpusz létezik, amely a hangzó szövegek mellett azok gépet leíratát, valamint annotációs szinteket is tartalmaz. Ez az átírási eljárás magas munkaerő- és költségigényével magyarázható.

Különösen csekély a magyar beszélt nyelvet reprezentáló korpuszok száma (Bodó és mtsai, 2017; Gósy, 2013). Közülük a Budapesti Egyetemi Kollégiumi Korpusz (BEKK) (Bodó és mtsai, 2017) esetében az interakciókat a résztvevők saját telefonjaikon rögzítették, ezért tulajdonképpen társalgásrészleteket tartalmaz. A BEszélt nyelvi Adatbázis (BEA) (Gósy, 2013) létrehozóinak fő célja az volt, hogy fonetikai (és nem szemantikai vagy pragmatikai) vizsgálatokat tegyen lehetővé, ezért a korpuszban alkalmazott annotációt is ennek megfelelően alakították ki. (A kutatócsoport egy újabb, hasonló korpuszról l.Szabó és mtsai (2021).) A fentebbi korpuszok tehát nem reprezentálják tökéletesen a teljes élőbeszédet, illetve feldolgozásuk módjával, annotációjukkal elsősorban fonetikai, és nem egyéb (pl. szemantikai vagy pragmatikai) vizsgálatokat kívánnak támogatni.

Mindezekkel összefüggésben a HuTongue hiánypótlónak tekinthető, hiszen nemzetközi szinten is rendkívül ritka a nagyméretű, informális beszélgetéseket tartalmazó korpusz. Tudomásunk szerint a HuTongue az első olyan magyar beszélt nyelvi korpusz, amely több száz órányi beszélgetésből áll és tartalmazza az összes beszélgetés részletesen annotált leíratát. Jelen dolgozatban a korpusz újabb, szűrt változatára támaszkodunk, melynek részletesebb bemutatását l.Vincze és mtsai (2021).

Mivel megfelelő magyar nyelvű adatbázisok hiányában eddig nem volt lehetőség a magyar beszélt nyelv beható korpuszalapú vizsgálatára, a jelen dolgozatban tett megállapítások értékes adalékul szolgálhatnak a magyar nyelvű beszélgetések néhány aspektusának vizsgálatához.

2. Kapcsolódó irodalom

Magunk a jelen dolgozattal azokhoz a kutatásokhoz szeretnénk hozzájárulást tenni, amelyek azokat a diskurzusokat vizsgálják, ahol egy adott csoport tagjai jelen nem levő személyekről beszélgetnek (részben tehát a pletykakutatásokhoz).

Az emberi informális kommunikáció jelentős részét, egyesek szerint kétharmadát, jelen nem levő, más személyekről folytatott értékelő tartalmú beszélgetés teszi ki (Dunbar, 1996, 2004). Jó pletykát hallani és pletykálni mindenki szeret, mégis magához a pletykához a köznapi értelemben pejoratív konnotációkat fűzünk, elítéljük azt. Miért létezik akkor, és miért olyan elterjedt a pletyka? Miért használunk ki szinte minden alkalmat arra, hogy másokat a hátuk mögött kibeszéljünk? (Galántai és mtsai, 2018).

Az erre vonatkozó ismeretanyag viszonylag szerény, és kvalitatív tapasztalatokra hagyatkozik. A pletyka mérését korábbi kutatások többnyire kvalitatív vagy kérdőíves kvantitatív módszerekkel (pl. Eckhaus és Ben-Hador (2019), írott szövegek alapján (Mittra és Gilbert, 2012) vagy megfigyelésen keresztül végezték (Dunbar, 2004).

A pletyka számos diszciplína által megközelített, ám eltérő mélységben vizsgált és különféleképpen meghatározott jelenség (Kisfalusi és Takács, 2018). Ezzel összefüggésben a pletyka fogalmának a szakirodalom alapján többféle megközelítése is ismert (Emler, 1994; Grosser és mtsai, 2012). A pletyka hagyományos definíciója szerint jelen nem levő személyről vagy személyekről folytatott értékelő tartalmat hordozó beszélgetéseket takar, melyekben egy értékelő tartalmat megfogalmazó személy és legalább egy hallgatója jelen van (Emler, 1994; Grosser és mtsai, 2012).

Míg a legtöbb tanulmány a pletyka egy vagy két céljára összpontosít, valójában – mint egy sokoldalú szociológiai jelenség – sok más társadalmi célja is lehetséges (Giardini és Wittek, 2019; Pápay, 2019). Azok az esetek például, amikor az üzenet feladója titkot árul el a pletyka vevőjének a pletyka tárgyáról, a diskurzus a feladó és a vevő közötti kapcsolat erősödhet, miközben a tárgy reputációja romlik. Ezzel összefüggésben, bár a pletyka fogalma gyakran negatív konnotációt hordoz, azonban fontos szerepet tölthet be az információáramlásban, a személyek közti kapcsolatok megerősítésében, a csoportnormák fenntartásában és betartatásában vagy szelepként szolgálhat a felgyülemlett negatív érzelmek „kiengedésében” (Grosser és mtsai, 2012).

Ami a HuTongue korpusz eddigi elemzési eredményeit illeti, egy korábbi dolgozatban a korpusz topikmodellezése történt meg az annotációval összefüggésben (Pápay és mtsai, 2018). A kutatók azt vizsgálták, hogy a pletykataggal ellátott, valamint pletykaként nem annotált diskurzusok topikjaiban milyen eltérések mutatkoznak. Egy a csoport egy másik dolgozatában a pletyka és a nem pletyka szövegek szófaji eloszlási különbségeit vizsgálta, valamint azt, hogy milyen grammatikai konstrukciókkal (névmások és igei morfémák) utal a pletyka feladója a pletyka tárgyára az üzenetében (Pápay, 2019).

A jelen dolgozat főbb kutatási kérdései a következők: Egyrészt azt szeretnénk feltérképezni, hogy azokban a diskurzusokban, amikor a csoporttagok egymásról beszélnek, mennyire dominálnak a különböző érzelmek és értékelő tartalmak.

Ezt a kutatási kérdést azért is tekintjük fontosnak, mert rámutathat a diskurzus motivációs tényezőire. A fentebbivel összefüggésben szeretnénk megtudni, hogy az, amikor a diskurzusok résztvevői egymásról beszélnek, mennyire gyakran teszik ezt negatívan. Megállapítható-e, hogy az egymásról való kommunikációt (beleértve a pletykát) javarészt negatív megállapítások, értékelések alkotják. Harmadrészt, szeretnénk megismerni az egymásról beszélés és a csoportdinamika összefüggéseit. Vajon hogyan mutat rá az egymásról beszélés a társas kapcsolatok minőségére, szorosságára?

A fentebbiekkel kapcsolatosan azt feltételezzük, hogy a beszéd mennyiségével nem feltétlenül fog korrelálni a negatív tartalom mennyisége. Másképpen, lehetséges, hogy valaki az összes megnyilatkozásának mennyiségéhez képest sokat beszél valakiről, és ezeknek a megnyilatkozásoknak éppen a pozitív értékelő tartalma van túlsúlyban a további tartalmakhoz képest. Ezzel összhangban azt is feltételezzük, hogy az egymásról beszélés pusztán mennyisége, valamint a pozitív és a negatív értékelő tartalmak mennyisége alapján nem azonos kapcsolatháló rajzolódik ki.

3. A vizsgálati korpusz

Ebben a részben bemutatjuk röviden a vizsgálati adatbázist. A korpusz szövegei hétköznapi kommunikációs helyzetekben keletkeztek, egy külső ingerektől elzárt környezetben (Szabó és Galántai, 2017; Galántai és mtsai, 2018; Gulyás és mtsai, 2018). A magas minőségű hanganyagot egy szórakoztatóipari cég rögzítette egy televíziós műsor készítése céljából. A hang rögzítése napi 24 órában történt, a korpusz pedig összesen 8 egymást követő nap felvételeit tartalmazza. A felvételek készítése során a megfelelő hangminőség biztosítása érdekében minden érintett mikroportot viselt ¹.

A korpusz létrehozásának elsődleges célja az volt, hogy megfelelő kutatási anyagot nyújtson a pletyka elemzéséhez az emberi interakcióban. A HuTongue egy egyedülálló adatbázis: legjobb tudomásunk szerint ez az első olyan nagyméretű magyar beszédkorpusz, amelyet teljes egészében kézzel írtak le és annotáltak.

A leiratozással párhuzamos annotációs folyamat kardinális lépése volt azoknak a diskurzusoknak az annotálása, amelyben a résztvevők jelen nem levő személyről vagy személyekről beszéltek. A kutatás fő tárgya a pletykaszövegek vizsgálata, azonban a korpusz annotálásához a kutatók előzetesen egy tág definíciót határoztak meg a pletyka fogalmára vonatkozóan. A munkadefiníció szerint az annotátoroknak minden olyan megnyilatkozást jelölniük kellett, amelynek tárgya egy vagy több jelen nem levő személy. Amennyiben a diskurzusban megemlített, de jelen nem levő személy az annotátor számára beazonosítható volt, akkor ezt egy megfelelő annotációs címkével kellett feltüntetnie. A tág definíció

¹ Mivel így módon a szövegek keletkezési körülményei (a résztvevők motivációi, valamint az alkalmankénti külső irányítás) befolyásolhatták a beszélői megnyilatkozásokat, a korpuszt félig vagy részlegesen spontánnak kell tekintenünk (Szabó és Szvetelszky, 2019).

célja az volt, hogy a lehető legkevésbé befolyásolják az annotátorokat a munkájukban azok pletykával kapcsolatos előzetes képzetei. Ennek köszönhetően tehát – ahogyan azt a bevezetőben is említettük – egy olyan korpusz jött létre, amely lehetővé teszi az emberi kommunikáció e szegmensének valamely definíció által előre nem korlátozott vizsgálatát.

A dolgozat további részében az annotáció alapján tett megállapításainkat tárgyaljuk a korpusz szentiment- és emóciótartalmaival összefüggésben.

4. A vizsgálatok és eredményei

A diskurzusokat a korpuszban elhelyezett annotációs tagekre támaszkodva elemeztük, vagyis a részletesebb vizsgálatokban csakis azon szövegrészletek vettek részt, amelyet az annotátorok a korábbi, tág meghatározás alapján jelen nem levő harmadik személyről szóló megnyilatkozásként jelöltek meg. A fenti elvárásnak megfelelő szövegrészletek a teljes, 8 nap hangfelvételeit tartalmazó korpuszból (főleg annak mérete miatt) automatikus eszközökkel lettek kiválogatva.

Az 1. táblázat azt mutatja meg, hogy a tokenek számában mérve ki kiről mennyit beszélt a korpuszban. A korpusz anonimizálása céljából a neveket természetesen mindenhol egységesen más nevekkal helyettesítettük.

ki/kiről	András	Dani	Gabi	Kornél	Sanyi	Vanda
András		3644	7088	2016	2797	2417
Dani	5131		4058	1412	2997	3438
Gabi	15489	5889		6514	5350	7395
Kornél	1925	905	2931		3114	2313
Sanyi	20838	7965	7856	10438		6304
Vanda	5098	6234	5125	2609	4233	

1. táblázat. A szereplők egymásról használt szavainak száma.

Elemzésünk fő irányát a szentiment- és emóciótartalmak vizsgálata képezte. Szentiment alatt azokat a szemantikai tartalmakat értjük, amelyek valamilyen (pozitívan vagy negatívan) értékelő tartalmat hordoznak, míg az emóciók valamely érzelem (pl. öröm, bánat, düh stb.) létére utalnak a szövegben (Zhao és mtsai, 2016). A szentiment- és emóciótartalmakat szótárak segítségével azonosítottuk a szövegekben. A feldolgozáshoz két, általános célra készült szentimentszótárat, valamint egy emóciószótárat használtunk. A szentimentszótárak közül az egyik lista (Liu, 2012) angol nyelvű szentimentszótárának fordítása volt, a másik egy bővebb, jelentősebb kiegészítést, így szleng szavakat is tartalmazó szótár (Szabó, 2015). Az emócióelemzéshez (Szabó és mtsai, 2016) szólistáit alkalmaztuk.

A vizsgálatok során számos összefüggést megvizsgáltunk:

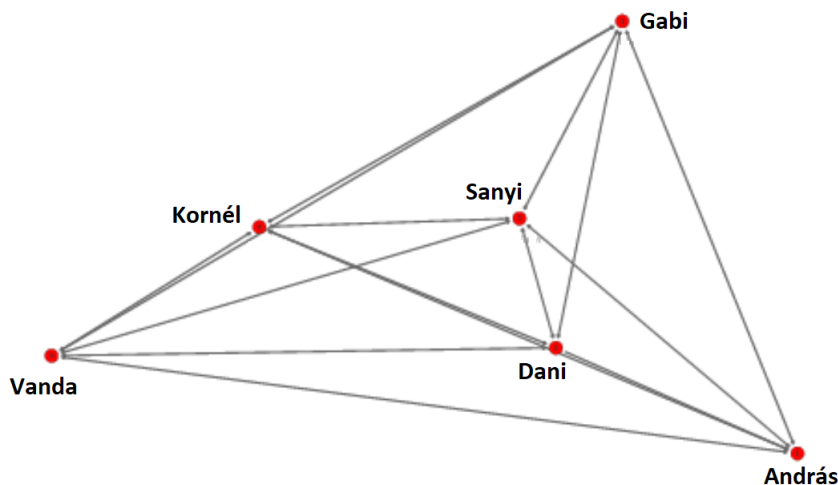
- az egyes emóciók és szentimentek gyakorisági előfordulása megszólalótól függetlenül,

- az egyes emóciók és szentimentek gyakorisági előfordulása megszólalónként,
- az egyes emóciók és szentimentek gyakorisági előfordulása megszólalónként a témaszemélyek szerinti bontásban,
- mindezeket kiszámoltuk úgy is, hogy az általunk vizsgált 8 emóciókategória pozitív és negatív elemeit összevontuk egymással (pl. öröm és szeretet, düh és bánat), és azokat a szentimentszótárakkal kapott eredményekkel összevontuk. Azt reméltük ugyanis, hogy ezzel a módszerrel pontosabb elemzési eredményeket érhetünk el, mintha pusztán a szentimentszótárakat használtuk volna fel.

Ezekből a jelen dolgozatban csupán néhány, érdekesebb összefüggés bemutatására vállalkozhattunk, ezért itt csak a pozitív és a negatív tartalmakra fókuszálunk (amelyeket a lista utolsó pontjának megfelelően az eredmények összevonásával nyertünk ki).

A bemutatott ábrákon a hálózati pontokat az egyes résztvevők adják, míg az éleket az, hogy az adott résztvevőről mennyit beszélnek az összes tokenszám viszonylatában, valamint, hogy az adott résztvevő mennyit beszél a többi résztvevőről az ő összes tokenszámaéhoz képest. Egy-egy résztvevőpár közötti él hosszát tehát a kettejük közötti kapcsolat két iránya együtt adja.

Első lépésben megnéztük tehát, hogy milyen kapcsolatháló rajzolódik ki kizárólag az egymásról beszélés mennyisége alapján. Itt tehát szemantikai tartalmakat még nem vizsgáltunk. Az így kapott hálózatot az 1. ábra mutatja be.



1. ábra: A szereplők szociális távolsága az egymásról használt szavak tükrében.

Ami az ábrán talán a legszembeötlőbb, az az, hogy Sanyi és Dani a vizsgált szempont szerint központi szerepet foglal el a hálózatban: a rövidebb élek

alapján róluk is sokat beszélnek, valamint ők is viszonylagosan sokat beszélnek másokról, és e további résztvevők és közöttük ez a mennyiség nagyjából arányosan oszlik meg. Az is látható, hogy ez a két résztvevő egymásról kiemelten sokat beszél. Hozzájuk képest például Vandának vagy Andrásnak nincs központi szerepe, csupán egy–két emberrel van szorosabb ilyen jellegű kapcsolatuk.

Következő lépésben megnéztük, az egymásról szóló beszéd tartalmak hány százaléka tartalmaz értékelő megnyilatkozást, azaz mekkora arányban találhatók bennük pozitív, illetve negatív töltetű szavak. Az eredményeket a 2. és 3. táblázatok mutatják.

Ki/Kiről	András	Dani	Gabi	Kornél	Sanyi	Vanda
András		21,90	23,38	20,44	25,28	17,91
Dani	22,88		23,63	17,49	22,86	20,77
Gabi	26,48	22,13		23,86	22,95	22,07
Kornél	19,01	10,39	21,08		19,52	17,42
Sanyi	20,84	17,31	18,79	18,68		17,20
Vanda	19,52	19,02	18,56	18,59	15,57	

2. táblázat. A résztvevők egymásról használt negatív szavainak aránya az összes egymásról használt szóhoz viszonyítva.

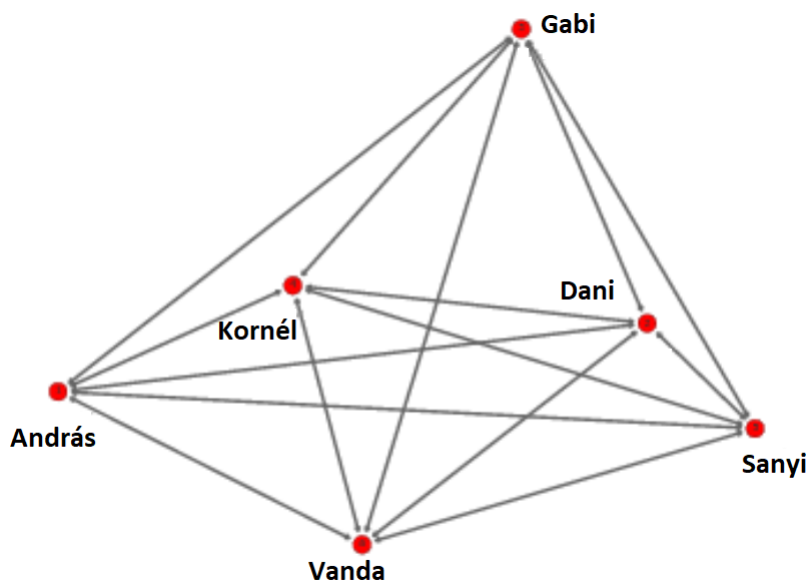
Ki/Kiről	András	Dani	Gabi	Kornél	Sanyi	Vanda
András		34,77	18,34	19,64	14,80	17,91
Dani	19,55		15,72	17,14	13,55	18,09
Gabi	17,41	19,58		19,36	19,96	21,54
Kornél	17,61	13,59	14,60		14,07	13,92
Sanyi	15,35	15,46	19,14	17,26		20,65
Vanda	15,06	14,90	21,62	16,33	18,99	

3. táblázat. A résztvevők egymásról használt pozitív szavainak aránya az összes egymásról használt szóhoz viszonyítva.

Az eredmények azt mutatják, hogy a csoport tagjai kimagasló számban használtak egymásról értékelő megnyilvánulásokat. Egyes esetekben ez akár a szavak 50%-a is lehet, például András Daniról szóló szavainak több mint fele sorolható be valamilyen érzelmi tartalmat hordozó csoportba, ami mindenképpen alátámasztja kettejük szorosabb kapcsolatát.

Ezek után figyeljük meg a negatív tartalmak alapján kirajzolódó kapcsolati hálót!

Azt látjuk, hogy a hálózat jelentős részben átrendeződik, amennyiben csupán a negatív tartalmakra fókuszálunk. Gyakorlatilag az előző ábrához képest csupán Dani őrzi meg valamelyest a pozícióját, ő a negatív tartalmak tekintetében is viszonylagosan sok szoros kapcsolattal rendelkezik. Ami a többi résztvevőt il-



2. ábra: A szereplők szociális távolsága az egymásról használt negatív szavak tükrében.

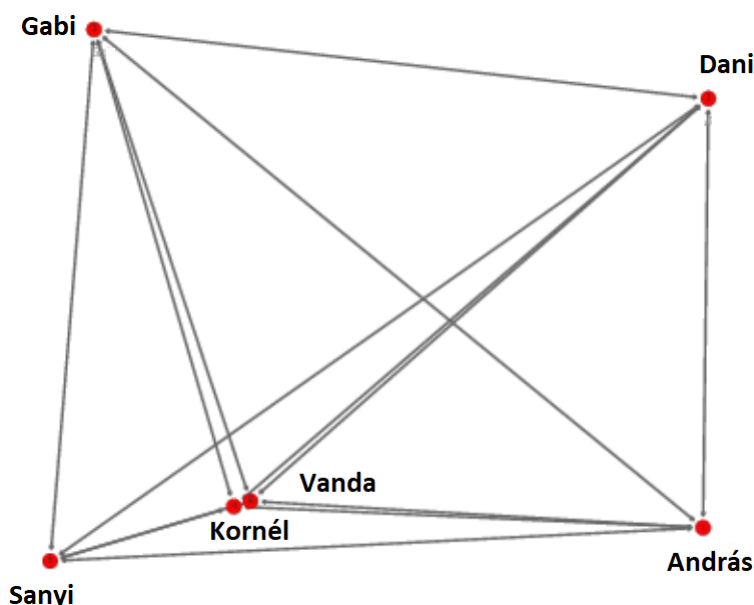
leti, Sanyi szerepe feltűnően csökken, Danival való szoros kapcsolata viszont a negatív kapcsolatrendszerben is megmarad. Érdekes viszont, hogy Vanda a negatív hálóban sokkal fontosabb szerepet tölt be, több szoros kapcsolattal a nem negatív háléhoz viszonyítva.

Végül, figyeljük meg a pozitív tartalmak alapján kirajzolódó kapcsolati hálót!

A pozitív kapcsolatháló felépítése feltűnően eltér mind a negatív, mind az általános hálózattól. Mindenekelőtt, a pozitív hálózati kapcsolatok jelentősen kiegyenlítetlenebbek mindkét vizsgált kapcsolatrendszerhez képest. Azt látjuk, hogy Kornél, Vanda és Sanyi kapcsolatai a legerősebbek egymással (közülük is leginkább az utóbbi két résztvevő), ugyanakkor másoktól nagyon távol esnek. Ha a többi résztvevőt nézzük, megállapítható, hogy azok kapcsolatai ebben a hálózatban rendre kifejezetten gyengék. Azt is látjuk, hogy amíg Sanyi és Dani pontjai a negatív hálóban nagyon közel voltak egymáshoz, a pozitív hálózatban távol esnek egymástól. Ez azt jelenti, hogy egymással kapcsolatos kommunikációjuk erősen negatív tartalmú.

5. Összegzés

A dolgozatban egy nagy méretű, magyar beszélt nyelvi adatbázis (HuTongue) segítségével azt vizsgáljuk, hogyan kommunikálnak egymásról a diskurzusok résztvevői, és megszólalásaikhoz milyen szentimentek és emóciók kapcsolódnak. Bár a



3. ábra: A szereplők szociális távolsága az egymásról használt pozitív szavak tükrében.

szakirodalom alapján az emberi informális kommunikáció jelentős részét, egyesek szerint kétharmadát, jelen nem lévő, más személyekről folytatott értékelő tartalmú beszélgetés teszi ki, az erre vonatkozó ismeretanyag viszonylag szerény, és kvalitatív tapasztalatokra hagyatkozik. Dolgozatunkkal ehhez a kutatási irányhoz kívántunk néhány korpuszalapú hálózatzvizsgálati eredménnyel hozzájárulni.

Összefoglalva a vizsgálati tapasztalatokat a következő legfontosabb megállapításokat tehetjük: Mindenekelőtt, az egymásról beszélés mennyisége, valamint értékelő szemantikai tartalma alapján egészen eltérő kapcsolati hálót lehetséges rajzolni. Mindez egybevág kezdeti hipotézisünkkel (1. 2. fejezet). Emellett lehetséges, hogy valaki (jelen esetben Sanyi) központi szerepet tölt be a hálózatban az alapján, hogy ő mennyit beszél másokról és róla mennyit beszélnek, azonban mind a pozitív, mind a negatív hálózatban inkább távolabbi kapcsolatokkal rendelkezik. Ez arra mutat, hogy az egymásról beszélés fontos csoportszervező funkciót tölthet be úgy is, hogy nem irányul a targetek negatív vagy pozitív értékelésére. A fenti tapasztalat ugyancsak egybevág a vizsgálatokat megelőző hipotézisünkkel. Azt feltételeztük ugyanis, hogy a beszéd mennyiségével nem feltétlenül fog korrelálni a negatív (és egyáltalán az értékelő) tartalom mennyisége. Másképpen, lehetséges például, hogy valaki az összes megnyilatkozásának mennyiségéhez képest sokat beszél valakiről, és ezeknek a megnyilatkozásoknak a pozitív értékelő tartalma túlsúlyban van a negatív tartalmakhoz képest. Az látszik tehát, hogy az egymásról beszéléssel nem nő szükségképpen a negatív érté-

kelés mennyisége. Mindebből a fordított irányban az is következik, hogy pusztán az egymásra vonatkozó beszéd mennyiségéből nem következtethetünk a vizsgált személyek társas kapcsolatainak tartalmi viszonyaira.

Köszönetnyilvánítás

A kutatást az Európai Kutatási Tanács (European Research Council), az Európai Unió Horizont 2020 kutatási és innovációs programjának keretében támogatta (ERC_CoG_2014_648693 sz. szerződésben), a kutatás vezetője Takács Károly.

Hivatkozások

- Bodó, Cs., Kocsis, Zs., Vargha, F.: A Budapesti Egyetemi Kollégiumi Korpusz. Elméleti és módszertani kérdések. In: Benő, A., Fazakas, N. (szerk.) Élőnyelvi kutatások és a dialektológia: Válogatás a 19. Élőnyelvi Konferencia - Marosvásárhely, 2016. szeptember 7-9. - előadásaiából. pp. 169–177 (2017)
- Crowdy, S.: Spoken corpus design. *Literary and Linguistic Computing* 8(4), 259–265 (1993)
- Dunbar, R.I.: *Grooming, Gossip and the Evolution of Language*. Harvard University Press, Cambridge, MA (1996)
- Dunbar, R.I.: Gossip in evolutionary perspective. *Review of General Psychology* 8(2), 100–110 (2004)
- Eckhaus, E., Ben-Hador, B.: Gossip and gender differences: a content analysis approach. *Journal of Gender Studies* 28(1), 97–108 (2019)
- Emler, N.: Gossip, reputation and social adaptation. In: Goodman, R.F., Ben-Ze'ev, A. (szerk.) *Good gossip*. University Press of Kansas, Lawrence (1994)
- Galántai, J., Pápay, B., Kubik, B.G., Szabó, M.K., Takács, K.: A pletyka a társas rend szolgálatában-az informális kommunikáció struktúrájának mélyebb megértéséért a computational social science eszközeivel. *Magyar Tudomány* 179(7), 964–976 (2018)
- Giardini, F., Wittek, R.: *The Oxford Handbook of Gossip and Reputation*. Oxford University Press (2019)
- Gósy, M.: Bea—a multifunctional hungarian spoken language database. *Phonetician* 105, 50–61 (2013)
- Grosser, T., Kidwell, V., Labianca, G.J.: Hearing it through the grapevine: Positive and negative workplace gossip. *Organizational Dynamics* 41, 52–61 (2012)
- Gulyás, A., Galántai, J., Szabó, M.K., Szebeni, Z.: A HuTongue spontán beszélt nyelvi korpusz leiratozásának és annotálásának minőségbiztosítási munkálatai. In: MSZNY 2018 - XIV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 317–330 (2018)
- Hemphill, C.T., Godfrey, J.J., Doddington, G.R.: The atis spoken language systems pilot corpus. In: *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990* (1990)
- Kisfalusi, D., Takács, K.: A pletyka és a reputáció összefüggései középiskolai osztályközösségekben. *Szociológiai Szemle* 28(1), 83–104 (2018)

- Liu, B.: Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5(1), 1–167 (2012)
- Maekawa, K., Koiso, H., Furui, S., Isahara, H.: Spontaneous speech corpus of japanese. In: *LREC*. pp. 947–9520. Citeseer (2000)
- Mengusoglu, E., Deroo, O.: Turkish lvcscr: Database preparation and language modeling for an agglutinative language. In: *IEEE International Conference on Acoustics Speech And Signal Processing*. vol. 6, pp. 4018–4018. IEEE; 1999 (2001)
- Mitra, T., Gilbert, E.: Have you heard?: How gossip flows through workplace email. In: *ICWSM* (2012)
- Neuberger, T., Gyarmathy, D., Grácsi, T.E., Horváth, V., Gósy, M., Beke, A.: Development of a large spontaneous speech database of agglutinative hungarian language. In: *International Conference on Text, Speech, and Dialogue*. pp. 424–431. Springer (2014)
- Oostdijk, N.: The spoken dutch corpus. overview and first evaluation. In: *LREC*. pp. 887–894. Athens, Greece (2000)
- Pápay, B.: The Purpose and Types of Organizational Gossip. *Doctoral Dissertation* (2019)
- Pápay, B., Kubik, B.Gy., Galántai, J.: Gossip is more than just story telling: Topic modeling and quantitative analysis on a spontaneous speech corpus. In: Jorge, A., Campos, R., Jatowt, A., Nunes, S. (szerk.) *Proceedings of the Text2StoryIR’18 Workshop, Grenoble, France, 26-March- 2018*. CEUR-WS.org (2018)
- Seppänen, T., Toivanen, J., Väyrynen, E.: Mediateam speech corpus: a first large finnish emotional speech database. In: *Proceedings of the Proceedings of XV International Conference of Phonetic Science*. pp. 2469–2472. Citeseer (2003)
- Szabó, M.K., Galántai, J.: Egy magyar nyelvű spontán beszélt nyelvi korpusz (HuTongue) létrehozásának tapasztalatai. In: *XXVI. MANYE Kongresszus konferenciakötete*. Pécs (2017)
- Szabó, M.K., Szvetelszky, Zs.: Részlegesen spontán körülmények között keletkezett pletykaszövegek pragmatikai szempontú vizsgálata. *Nyelvtudományi Közlemények* 115, 317–343 (2019)
- Szabó, M.K., Vincze, V., Ring, O., Üveges, I., Vít, E., Samu, F., Gulyás, A., Galántai, J., Szvetelszky, Zs., Bodor-Eranus, E.H., Takács, K.: StaffTalk: magyar nyelvű spontán beszélgetések korpusza. In: *XVII. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem, Szeged (2021)
- Szabó, M.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái (‘Experiences and dilemmas of the creation of a Hungarian sentiment dictionary’). *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához*. 177, 278–285 (2015)
- Szabó, M., Vincze, V., Morvay, G.: Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. Reményi, AÁ, Sárdi Cs., Tóth Zs.(szerk.) *Távlatok a mai magyar alkalmazott nyelvészetben*. Budapest: Tinta pp. 282–292 (2016)
- Van Bael, C., Baayen, R.H., Strik, H.: Segment deletion in spontaneous speech: a corpus study using mixed effects models with crossed random effects. In: *INTERSPEECH*. pp. 2741–2744 (2007)

Vincze, V., Üveges, I., Szabó, M.K., Takács, K.: A magyar beszélt és írott nyelv különböző korpuszainak morfológiai és szófaji vizsgálata. In: XVII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2021)

Zhao, J., Liu, K., Xu, L.: Sentiment analysis: mining opinions, sentiments, and emotions (2016)

Egy szabály- és következtetésvezérelt csevegőrobot modell anatómiája

Az ITSy-Bitsy modell

Kilián Imre

Deutsche Telekom Systems Solutions Hungary Kft.
7621 Pécs, Hungary
Imre-Zoltan.Kilian@t-systems.com
tel: +49 69 9731792500
www.deutschetelekomitsolutions.hu

Kivonat Csevegőrobotok az elmúlt években robbanásszerűen terjedtek el az ipari szoftverekben, megvalósítva egy természetes nyelvű csevegést lehetővé tevő, újszerű kezelői felületet. Az ipari környezetben használt csevegőrobotok neurális hálós/mélytanulási technológiával készülnek. Itt most mégis egy szabály- és következtetéses alapon készült csevegőrobot modellről számolunk be, és részletezzük a nyelvi feldolgozás menetrendjét, különös tekintettel a tartalmi elemző és átalakító modulra. A cikk végén összehasonlítást teszünk a mélytanulási technológiával elérhető képességek között.

1 Bevezetés

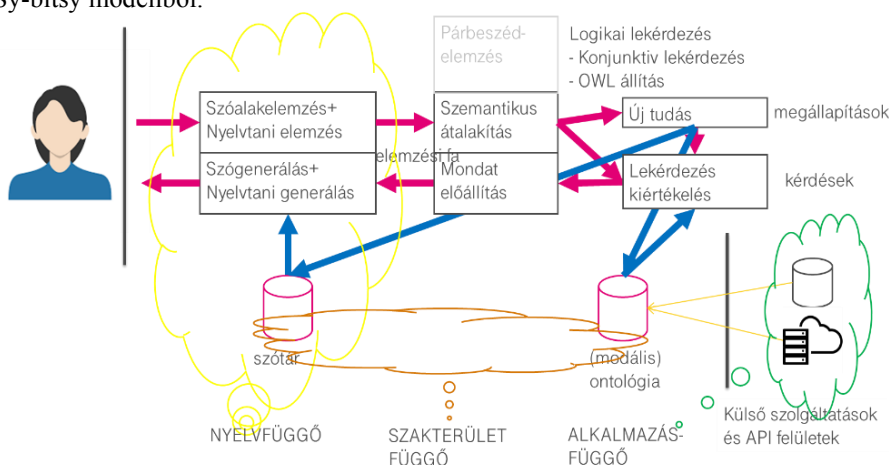
A *csevegőrobotok* (*chatbotok*) nem kevesebbet, mint az Alain Turing kitűzte célt követik: olyan gépi eszközöket valósítanak meg, amelyek természetes nyelvű párbeszédre (csevegésre) keresztül érnek el valamilyen célt – amit a háttérben működő alkalmazói program valósít meg (Corydon 2012). A modern robotok képességei ezért általában valamiféle kötött tématerületre vannak szabva. A csevegőrobot programozás alapvető gyakorlata pl. egy Gyakran Ismételt Kérdések (GYIK) tudásanyag betanítása.

Az 1966-ban Joseph Weizenbaum által készített ELIZA robot (Weizenbaum 1966) az első efféle volt. A robot egy pszichiáter szerepét játszotta, és a páciens begépelte mondataiból szövegszerű átalakításokkal kérdéseket csinált úgy, hogy az átalakítási mintáiból véletlenszerűen választott egyet. Az angol nyelv különösen alkalmas az „I am” → „You are” jellegű átalakításokra. Például, ha a páciens a következő mondatot gépelte be: „I am deprived today.”, arra a következő választ (vagy valami hasonlót) adhatott: „Why are you deprived today?”

Az iparban használt csevegőrobotok jellemző megoldása: nagyobb cégeknek saját csevegőmotorjuk van, amit neurális háló valósít meg. Ezek többcélúan testre szabhatók, a szókészletük, a felismert mondatmintáik, sőt, a párbeszéd mintáik programozhatók, a felismerést végző hálózat betanítható, és az adatbázisokba vagy szolgáltatásokba történő lekérdezésük pedig konfigurálható. (Watson, Rasa). Az alább leírt megoldás az

ITSy-bitsy, szabály- és következtetésvezérelt csevegőrobot felépítését tárgyalja, amely Prolog programkörnyezetben született.

A nyelvfeldolgozás klasszikus menetrendjét Vauquois óta ismerjük [4], az ITSy-bitsy szintén ezt az elvet követi. A feldolgozás a bemenő szöveg szimbólumokra tördelésével kezdődik, amit a szóalakelemző egy absztrakt szóalak-szerkezetté alakít át. Ez a nyelvtani elemző bemenete, amely a mondatból egy elemzési fát hoz létre. A nyelvtani elemzés feletti réteg a lehet a párbeszédelemzés, ilyen lehetőséget az ITSy-bitsy jelenleg nem nyújt. A mondatok (bekezdések) elemzési fájából indul ki a tartalmi elemzés, ami egy kiértékelésre, ill. végrehajtásra kész logikai alakot bocsát ki magából. Az ezen réteg feletti pragmatikus szint a szándékok és lehetőségek, netán metakommunikációs eszközök figyelembevételével árnyalja a megértést. Ez a réteg szintén hiányzik az ITSy-bitsy modellből.



1. ábra Az ITSy-bitsy felépítése a klasszikus elveket követi

2 Nyelvi szint

2.1 Nyelvi elemzés

A nyelvi elemző angol nyelvre lett megvalósítva, amelyben a szóalakelemzés feladata leegyszerűsödik, és a kötött szórend a nyelvtani elemzést is megkönnyíti.

A szóalakelemzés persze nem tűnik teljesen el. Mivel formailag az *s* igerag a többes számtól nem különíthető el, és több szó egyidejűleg többféle szófajként is elemezhető, már a szóalakelemzőnek is *nemdeterminisztikusan* kell működnie. (Pl. a 'works' szóalak egyaránt elemezhető igeiként és többes számú főnévként.)

A nyelvtani elemző alulról-felfelé irányban, szintén nemdeterminisztikusan működik, lényegileg a Cock-Younger-Kasami elemzőalgoritmus Prologra alkalmazott, *mélyégi, visszalépéses keresési stratégiával* megvalósított változata. Az elemző

Contralogra épül, ami a Prologgal illeszkedő, ahhoz készített, előrehaladó módon következtető előfordító (Kilián, 2016).

Az alulról felfelé működésmód miatt az elemző a szavak és kifejezések összes lehetséges elemzését előállítja. Ezek közül egyesek a további elemzési lépések során elhalnak, mások viszont részt vesznek a mondat teljes elemzési fájának felépítésében. Ennek során két gond merülhet fel. Egyrészt az elhaló elemzési ágak előállítása felesleges, de ezt korábban még nem mindig tudjuk eldönteni. Másrészt az elemző alternatív elemzések tömegét állíthatja elő, amiről esetleg nem is tudjuk, hogy melyiket fogadjuk el belőlük érvényesnek. Persze léteznek tényleg többféleképpen elemezhető (és értelmezhető) mondatok is, ilyenkor a többszörösség kezelése (nemdeterminizmus) nemigen megtakarítható.

Az elemzések előállítása során az alternatív elemzések nem látnak át egymás részeredményeire, ezért nem tudunk különféle heurisztikus kikötéseket tenni. (pl., hogy egy adott fa-mintára illeszkedően a lehető legnagyobb szövegszeletet fogadjuk el, ill. hogy egy adott szövegszeletre illeszkedő lehető legegyszerűbb fa legyen a nyerő.)

Nyelvtani elemzőnk kimenete az elemzett mondat *elemzési fája*. Ez egyrészt a Prolog megvalósítási nyelven természetes módon ábrázolható, másrészt az elemzési fából *szöveget generáló* modul nem csinál mást, mint az elemzési fát járja körül.

Az ITSy-bitsy a jelen állapotában az A2 (erős kezdő) szintre van beprogramozva. A szótárszerkezet bővíthető, történt már sikeres kísérlet a B2 (erős haladó) szint elérésére, mindenesetre nyelvtani bővítés nélkül. A szótárbővítés lehetővé teszi szakterület-függő szókészlet beépítését is. Végso célként a C (anyanyelvi) szint is kitűzhető, de félo, hogy a szaporodó intuitív megértési problémák miatt nemigen érhető el.

2.2 A nyelv határai

Az elemzési alapfeladat egyes – jól formált – mondatokat elfogad, másokat elvet. Készíthető olyan elemző, amely az akadémiai nyelvet elfogadja, a többit viszont elveti. Elvetheti pl. egyes területileg, szubkultúrák, életkorok, netán foglalkozások által meghatározott nyelvjárások mondatait. Másrészt viszont akadémiai nyelv nem létezik, mert azt a legjobb esetben is csak írásban használjuk – még a legszabatosabb beszélő mondatainak is csak kis része fér bele az akadémia keretekbe. Sejtésünk, hogy az akadémia nyelvet bármilyen tökéletesen elemezze egy program, az már a ténylegesen használt írott nyelv vizsgáján is megbukik, a beszélt nyelvről nem is beszélve.

3 Tartalmi elemzés

A tartalmi elemzések alapfeltételezése, hogy a mondatok logikai alakját a háttérben, *fesztes logikai alakban tárolt, és hétköznapi fogalmakat tartalmazó tudásállományhoz* illesztjük, és megállapítjuk, hogy következhet-e abból, netán ellentmondásban áll-e vele. A háttérben tárolt hétköznapi tudást régebben *tudásbázisnak*, manapság inkább *ontológiának* nevezik.

3.1 Statikus szemantika

Nyelvtanilag jól formált mondatokkal is lehet értelmetlen dolgokat mondani (...mert mi lenne akkor a költőkkel? /Bach Iván/). Ezt az ontológia modellszintje alapján az elemzett mondatok *statikus szemantikus* ellenőrzésével ellenőrizhetjük, ami egyes felesleges elemzési tévutak szűrésére is használható. (pl. „Színtelen zöld eszmék dühödten alsznak.” /Chomsky/)

Ha ugyanis a modellszinten az egyes relációk értelmezési tartományát is tároljuk, akkor a fenti mondat pl. a következő szemantikai szabályokat sérti:

$\text{Dom}(\text{Verb}(\text{sleep})) \cap \text{Concept}(\text{idea}) = \emptyset$ az 'aludni' ige értelmezési tartományának és az 'eszme' fogalomnak nincs közös része.

$\text{Dom}(\text{Adj}(\text{green})) \cap \text{Concept}(\text{idea}) = \emptyset$ a 'zöld' jelző értelmezési tartományának és az 'eszme' fogalomnak nincs közös része.

$\text{colorless} \cap \text{colorful} \supset \text{green} = \emptyset$ a 'színes' és a 'színtelen' egymás kiegészítő halmazai, 'zöld' részhalmazzal nincs közös része.

3.2 Átalakítás logikai nyelvvé

Az átalakítás során a mondat elemzési fájából egy *logikai nyelvű mondatot* hozunk létre. Alapvető kérdés a logikai nyelv meghatározása: Nagyon magas szintű logikai nyelvnek nagy a kifejező ereje, ám ezek a nyelvek legtöbbször *eldönthetetlenek*. Alacsonyabb szintű nyelvek annyira szószátyárok, hogy a gyakorlatban sajnos használhatatlanok is. Emellett fontos szempont lehet a nyelvhez készített megoldó/kiértékelő szoftver csomag elérhetősége. Ebből a szempontból a Prolog ideálisnak tűnik.

A logikai alakba történő átalakítás nagymértékben függ a mondatban használt *igemódtól*.

A *feltételes módtól* ezúttal eltekintünk, mert az az angol nyelven csupán feltételes kötőszóval összekapcsolt, és időbeli viszonyba állított kijelentő tagmondatokkal van megvalósítva.

Felszólító mondatokat a csevegőrobotok valami azonnali tevékenységre utaló parancsként értelmezhetnek. Közvetlen tevékenység végrehajtásának különböző ügyfélszolgálatok esetén sok jelentősége lehet, de az irodai munka is tartogat efféle lehetőségeket.

Különösen érdekes a *kijelentő és a kérdő mondatok* megvalósítása. A kérdő mondatokat jelenleg az elsőrendű logikai modellben ún. *konjunktív kérdésekké*, vagyis elemi relációk konjunkciójává alakíthatjuk át, amelyet a Prolog alapú tudáskezelő rendszer értékel ki és válaszol meg.

A kijelentő mondatokat viszont aszimmetrikus módon OWL (Web Ontology Language) állításokká képezzük le, majd a tudásállományhoz hozzávesszük.

3.3 Kétszintű logikai modell

A hétköznapi élet mondatai és kérdései sajnos nem mindig elsőrendűek. Az efféle problémák kezelését ún. *kétszintű tudásábrázolással* oldhatjuk meg, amit *reifikációnak* is neveznek. A *példányszintű / adatszintű* tudáselemeken (pl. Kala Pál mikor hol, mit csinált) kívül *modellszintű* tudáselemeket (tudásszegmenst) is tartalmaz, ami a példányszinten használt modelleszközökre: osztályokra, tulajdonságokra és relációkra vonatkozó általános ismereteket, pl. az értelmezési tartományait, ill. azok egyéb összefüggéseit tárolja.

A kétszintű tudásábrázolás már másodrendű kérdésekre is választ adhat. Pl. a „Mi a különbség Malacka és Bagoj (sic!) között?” kérdésből egy olyan logikai kifejezést hozhatunk létre, amely egyrészt modellszinten megkeresheti a két állatkára csak külön-külön vonatkozó relációkat, de a közös tulajdonságok és relációk különböző értékeit is megkeresheti. Például a *hasKeeper/vanGazdája* tulajdonság értéke mindkettőjükre „Christopher Robin”, míg a *hasResidence* reláció értéke különböző, a *nestingTime/költésiIdő* reláció viszont Malackára, mint emlőslátra nincs is értelmezve.

3.4 Modális logikai modell

Logikai alapú rendszerek egyik gondja az *ellentmondás-mentesség*. Effélék nem modellezhetők, pedig a valóságos élet tele van ellentmondásokkal. Ellentmondások kezelésének egyik útja a *modális logikák* bevezetése, ahol *világocskákat (környezeteket)* határozunk meg. Az ellentmondás-mentesség csak egy világocskán belül követelmény, a világok között már nem.

A világocskák persze lehetnek *halmazok* is. A jogi normatívák pl. az ún. *deontikus* világocská-halmazban értelmezhetők, amely a TILOS, MEGENGEDETT, VÁLASZTHATÓ, KÖTELEZŐ világocskákat (operátorokat) tartalmazza. Az egyes világocskák között a következtetések végzéséhez különféle logikai axiómákat szokásos megállapítani.

Világocskákat az *episztemikus-doxasztikus* logika alapján *hozzárendelhetünk gondolkodó ügynökökhöz (ágensekhez)* is, pl. *Know*(WinnieThePooh), *Believe*(Piglet). A világocskák egymásba is ágyazhatók, pl. *B*(Piglet, *K*(Heffalump)). Az egymásba ágyazott világocskák között gyakori az *öröklés*, mint logikai axióma használata, a világocskák globális gyökéreleme ilyenkor az általánosan elfogadott/megdönthetetlen állításokat tartalmazó világ (Alberti 2009, Kilián 2012).

A világocskákat operátorként értelmezve a logikai kifejezések elé írjuk, pl: *K*(WinnieThePooh) (*HasLittleBrain*(Ego) , *Bear*(Ego) , *Child*(Ego))

... vagyis: Micimackó tudja magáról, hogy ő egy csekélyértelmű medvebocs.

A mondatok időparaméterét – mint a mondat egészére vonatkozó információt – szintén felfoghatjuk modális környezetként, amit külön dimenzióként együtt alkalmazhatunk az egyéb világocskákkal. Személetes ábrázolásmódban az egyéb világocskák „szőreiként”, vagyis beágyazott apró *levél-világocskákként* érdemes őket ábrázolni, és a következtetések elvégzésére itt is rögzíthetők különféle axiómák (Ruzsa, 1984) Időlogikában a következő operátorok használatosak:

BeforeSometime(T) a T időpont előtt valamikor igaz volt

$\text{BeforeAlways}(T)$	a T időpont előtt mindig igaz volt
$\text{AfterSometime}(T)$	a T időpont után valamikor igaz (lesz)
$\text{AfterAlways}(T)$	a T időpont után mindig igaz (lesz)

Példa axiómákra, amelyek segítségével logikai következtetések végezhetők:

$W_{\text{Off}} \phi \leftarrow W_{\text{Anc}} \phi$	a világocskák közötti tudásöröklődés axiómája
$K(WHO) \phi \rightarrow B(WHO) \phi$	episztemikus fokozatöröklődés axiómája
$AS\phi \vee AS\psi \rightarrow AS(\phi \vee AS\psi)$	
$\vee AS(AS\phi \vee \psi)$	
$\vee AS(AS\phi \vee AS\psi)$	ha a jövőben valamikor egy dolog vagy egy másik igaz lesz, akkor vagy az egyik igaz és a másik még nem, vagy fordítva, vagy még egyik sem igaz

3.5 Átalakítási szabályok

Az alábbiakban állítások és kérdések konjunktív logikai lekérdezőnyelvvé történő átalakítását mutatjuk be elsődrendű logika felett.

Igei kijelentő mondat modellje elsődrendű nyelven:

$\text{alany}(\dots), \text{argumentumok}(\dots) \rightarrow \text{ige}(\dots)$

Névszói kijelentő mondatok esetében az állítmány nem ige, hanem névszó, tehát megfelelő logikai kifejezést is a névszói szerkezetekhez hasonlóan képezzük. *Névszói szerkezetek* átalakítási mintája az ún. *összegző (kumulatív)* értelmezés szerint:

$\text{jelző}_1(X), \dots, \text{jelző}_N(X), \text{köznév}(X)$

Tulajdonnevek egy egyedi azonosítóba képződnek le. Az azonosító az őt tartalmazó legszűkebb osztályon belül egyértelmű, vagyis az azonosítót az osztálynévvel is címkézzük. Pl. a „Malacka a Százholdas Pagonyban lakik.” állításból az alábbi logikai (tény) -állítás keletkezik.

$\text{live}(\text{piglet}\#\text{pig}, \text{hundredAcreWood}\#\text{woodland}).$

Az általános modellhez képest természetesen kivételek is megfogalmazhatók, pl. a „szétszórt bölcsészlány” kifejezés inkább önmagában is egy következtetés ($\text{szétszórt}(X) \leftarrow \text{lány}(X), \text{bölcsész}(X)$) (Alberti, 2011) Emellett az igei kezelési módja sem mindig egyöntetű.

- *Tranzitív ige*k bináris, *ditranzitív ige*k három oldalú relációkba fordíthatók, és ezek sorrendje rögzített ugyanúgy, mint a kötelező vonzatoké is.
- Egy-egy *ige egyéb paramétereit* azonban vonzat helyett gyakran *szabad határozókként* ismeri fel az elemző. Az ontológia felépítésétől függően ezeket gyakran újabb relációparaméterként kell beszúrni.
- Egyes *általános jelentésű ige*k relációs fordítását újabb *típus paraméterrel* is kiegészítjük. Tipikusan ilyen a *have* ige, amelyet ilyen módon a birtoktárgy típusától függően részrelációkra bontunk.

Kijelentő mondatok átalakítása tényállításokat eredményez. Ha az alany és az esetleges igei argumentumok leírása bonyolultabb (pl. kvantorokat, birtokviszonyt stb. is tartalmaz), akkor a logikai alak egy Horn-klóz, vagyis egy Prolog szabály lesz. Attól függően, hogy az új tényállítás vagy szabály már létezik-e, döntünk annak felvételéről. További kérdés, hogy *melyik világocskába* vegyük fel az új ismeretet, amit stratégiai megfontolásokkal dönthetünk el. Egy *bizalmatlan* stratégia (pl. egy bíró a bíróságon) minden információt a kijelentést tevő egyén világocskájába helyez, míg egy *hiszékeny* stratégia mindent a gyökérvilágba, vagyis a megdönthetetlen állítások közé.

Eldöntendő (yes/no) kérdő mondatok esetében a mondat logikai alakját kiértékeljük, és a logikai eredményt visszaadjuk.

Kiegészítendő kérdésekben a kérdőszó vagy kifejezés gyakran egy igei vonzat szerepét játssza. A kérdőszóból logikai változó lesz, az egész mondat pedig olyan logikai kifejezéssé fordul le, amely ezt tartalmazza. A kérdés kiértékelése során a változó értéket kap, ez lesz a kérdésre adandó válasz. A kiegészítendő kérdés modellje tehát:

$X^{\text{mondat}}(X)$.

Kvantorok az elemzett nyelvben előfordulhatnak explicit módon is, de több ösztönös vagy rejtett nyelvi kifejezés is gyakran csak kvantorokkal (*gyűjtőfüggvényekkel*) értelmezhető. Jelenleg a következő gyűjtőfüggvények kezelése van megvalósítva.

- Egyes számú, de a nem konkrét főnévi csoport esetén a kérdésből létrehozott egyszerű kérdés kerül végrehajtásra. Ha az ismeretlennek több lehetséges értéke is volna, akkor visszalépéssel előállítja őket, pl. a „Who is Baby Roo’s male friend?” kérdésre válaszolva.

```
X:=X^(male(X), friend(X),
      have(babyRoo#kangaroo, friend, X))
```

- Többes számú főnévi csoport esetén az eredmény egy halmaz. Az ilyen kérdésből létrehozott lekérdezés (ld. alább) a COND feltételnek megfelelő X változókat gyűjti össze a LIST változóba, pl. a „Who are the male friends of Piglet?” feltevése esetén.

```
LIST:=each(X^male(X), friend(X),
           have(babyRoo#kangaroo, friend, X))
```

- Ha határozott névelőt használunk, akkor arra utalunk, hogy a körülírt dologból egyetlenegy létezik. Az ilyen kérdésből létrehozott lekérdezés az *egyetlen, a feltételeknek megfelelő változóértéket* adja vissza.

```
VALUE:=only(X^COND)
```

- Gyakran valamilyen a feltételnek megfelelő *gyűjtemény számosságára* kérdezzünk. Ez az alábbi lekérdezéssel lehetséges:

```
NR:=count(X^COND)
```

- *Helyhatározói kérdésekre* van egy további érdekes beépített feldolgozás. Ha ugyanis több szereplőre vonatkozólag valami helyhatározós kérdés tennénk fel, (pl., hogy hol laknak), akkor a földrajzi egész-rész reláción (hasGeopart/2) keresztül a legközelebbi közös lakóhelyet kapjuk válaszul (lca: least common ancestor). A kérdés általános alakjában paraméterül adjuk a relációt magát is, ld. alább.

$X := \text{lca}(X^{\text{CALL}}, \text{RELATION})$.

Például, a „Where do Piglet and Owl live?” kérdésre a „HundredAcreWood” választ kapjuk, míg a „Where do Kanga and Baby Roo live?” kérdésre a kangasHouse a helyes válasz. (Az eltérő írásmód azt jelzi: egy azonosítóról, és nem valami-féle kifejezésről van szó, amelynek nyelvtani kifejtése nincs pontosabban meghatározva.)

3.6 Kapcsolat konkrét szolgáltatásokhoz

A szoftver tartalmaz egy csatolómodult, amelyben a konkrét Prolog ontológiához történő kapcsolódás van leírva: vagyis esetleges Prolog átalakító-szabályokat lehet/kell itt megadni. Külső adatbázisok vagy szolgáltatások esetén meg kell valósítani az ontológiát kiterjesztő műveleteket is (vagyis ami lekérdez, bővít, töröl egy olyan tudáselemet, amely külső eszközön vagy szoftveren áll rendelkezésre).

4 Összehasonlítás mélytanulásos megoldásokkal

A napjainkban elterjedt mélytanulásos megoldások sikere kikényszeríti, hogy egy szabályalapú megoldás készítője legalábbis áttekintse, kiértékelje egy mélytanulásos eszköz működését, és – ha nem is pontos mérési módszereken keresztül – de legalább hevenyészett összehasonlítást tegyen.

Heading level	Szabályvezérelt	Mélytanulásos
Tanítás	kevesebb szabállyal	sok példával
Működés	hajszálpontos	valószínűségi
Be nem tanított esetek	intoleráns	toleráns
Következtetés	van	korlátozott
Metaszintek	tudja	nem tudja
	hasonlít az anyanyelv-használathoz	hasonlít az idegen nyelv használatához

2. ábra Szabályvezérelt és a mélytanulásos megoldások összehasonlítása

Az összehasonlításból végeredményben az derül ki, hogy a következtetésvezérelt megoldás az általa szállított megoldásokra hajszálpontos, viszont be nem tanított megoldásokra intoleráns. Talán kicsit az anyanyelvi beszélő nyelvértéséhez hasonlít, aki általában a közölt mondat minden egyes szavát pontosan érti, és pontosan meg tudja

válaszolni, de esetleg egy akár tájnyelven használt, nem ismert vagy oda nem illő kötő- vagy indulatszó komolyan megzavarhatja a megértést.

A mélytanulós megoldás talán inkább az idegen nyelvek megértését mintázzák. Ha egy-két szó ismeretlen is egy mondatban, attól még a lényeget megértjük, és közben persze bizakodunk, hogy a fel nem ismert szavak nem változtatják meg gyökeresen a mondat értelmét.

Talán az intelligens beszélő stratégiája a legszerencsésebb: ha pl. egy csángó nyelv-járású adatközlő minden szavát nem érti hajszálpontosan, stratégiát vált, és valószínűségi alapon, valamint egyes metakommunikációs jelekből fogja a mondanivalót. Az efféle megoldásokat a műszaki életben *hibridnek* nevezik.

5 Értékelés és további munkák

Az ITSy-bitsy a jelen állapotában egy állapotmentes csevegőrobot, amely angol nyelven működik. A moduláris felépítésből következőleg más elemző vagy logikai alakra hozó modul is bekapcsolható, és más logikában másféle kiértékelés is használható. Annak sincs akadálya, hogy valamelyik modult másféle technológiára (pl. mélytanulósosra) cseréljük (Wei-Rui, 2018).

A robot az SWI-Prolog rendszer környezetében készült (Wielemaker, 2003). A rendszerrel lényegileg végigjártuk a csevegőrobot készítés főutcáját. Létrejött egy műszakilag teljes értékű modell, és minden fontos technológiai lépésre kiterjedő feldolgozási lánc, ami jelenleg a kiegészítendő kérdések feldolgozását tudja a legtokéletesebben elvégezni (és megválaszolni).

A cikkben leírt kezdeményezés ígéretes, de a további életét teljesen biztosan befolyásolja a *külső érdeklődés*: az a segítség, amivel legalább egy életszerű és életszagú mintaalkalmazás összehozható. Természetesen – magyar környezetben – eléggé fontos volna a magyar elemző létrehozása és beillesztése is...

Hivatkozások

- Alberti, G. 2009. ReALIS: An Interpretation System which is Reciprocal and Lifelong. Workshop 'Focus on Discourse and Context-Dependence' UvA, Amsterdam Center for Language and Comm., 2009.
- Alberti, G.: ReALIS: Interpretálók a világban, világok az interpretálóban. Akadémiai Kiadó, Budapest (2011)
- Corydon, I.: Alain Turing at 100, The Harvard Gazette, Harvard University, September, 2012, (<https://news.harvard.edu/gazette/story/2012/09/alan-turing-at-100>, elérés: 21-07-2020)
- Kilián, I.: ReALIS: egy többszereplős, episztemikus rendszer Prolog modellje. In: SzámOkt 2012. konferencia kiadványa, pp. 276–281, Erdélyi Magyar Műszaki Tudományos Társaság, Kolozsvár (2012)
- Kilián, I.: Contralog: a Prolog conform forward-chaining environment and its application for dynamic programming and natural language parsing, In: Acta Universitatis Sapientiae, 8-1, pp.41-62. Marosvásárhely, 2016.
- Ruzsa, I.: Klasszikus, modális és intenzionális logika. Akadémiai Kiadó, Budapest (1984)

- Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In: IFIP Congress (2) pp. 1114-1122
- Wei W., Rui, Y.: Deep Chit-Chat: Deep Learning for Chatbots. Conference on Empirical Methods in Natural Language Processing, Brussels, (2018) <http://www.ruiyan.me/pubs/tutorial-emnlp18.pdf>, elérés: 05-12-2020)
- Wielemaker J.: An overview of the SWI-Prolog programming environment. In: Proc. 13-th International Workshop on Logic Programming Environments, pp.1-16. eds: F. Mesnard, A. Serebenik, Katholieke Universiteit, Leuven, Belgium (2003)
- Weizenbaum, J.: ELIZA--A Computer Program for the Study of Natural Language Communication Between Man and Machine (http://www.universelle-automation.de/1966_Boston.pdf, elérés: 21-07-2020).

A gépi elemzők kriminalisztikai szempontú felhasználásának lehetőségei

Vincze Veronika¹, Kicsi András^{1,2}, Főző Eszter³, Vidács László^{1,2}

¹MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szeged, Tisza Lajos körút 103.

²Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
Szeged, Dugonics tér 13.

³Nemzetbiztonsági Szakszolgálat
Budapest, Törökvész út 32-34.
{vinczev,akicsi,lac}@inf.u-szeged.hu
fozo.eszter@nbsz.gov.hu

Kivonat A modern kommunikációs csatornák használatával a korábbanál jelentősen könnyebbé vált üzenetek anonim módon való közlése közönség, vagy akár kiválasztott emberek számára is. Ez visszaélésekkel is jár, a különböző zsaroló, fenyegető vagy rágalmazó üzenetek forrása is titokban marad. Egyes esetekben a bűnügyi hatóság sem talál erre közvetlen bizonyítékot. A szöveg azonban szükségszerűen tartalmaz bizonyos, a szerzőre jellemző jegyeket. A szöveg stilometriai vizsgálata egy értékes eszköz ilyen nyomozati körülmények között. Ez megköveteli a stílusjegyek azonosítását, amelyet jelenleg a nyelvészeti szakértő a szöveg alapos tanulmányozásával tár fel. Ezt segítő, magyar nyelvű szövegek elemzésére is alkalmas automatizált megoldás nem áll rendelkezésre. Tanulmányunkban azt vizsgáljuk, hogy magyar nyelvű szövegtörzsek automatizáltan kinyert különböző stílusjegyei milyen mértékben utalnak a szerző személyazonosságára, ezzel a későbbiekben hozzájárulva a szakértő munkájához. Az elemzés során számos statisztikai, morfológiai, szintaktikai, szemantikai és pragmatikai jellemző értékeit vetjük össze négy szerző összesen 61 dokumentuma felett, melyek között bűnügyi írárok is találhatók. Eredményeink rávilágítanak, hogy a szövegek automatikusan azonosított stílusjegyei alapján lehetséges a szerzők pszichológiai jellemzése, ami a későbbiekben segítheti a bűnügyek felderítését.

Kulcsszavak: szerzőazonosítás, stilometria, kriminalisztika, NLP

1. Bevezetés

Habár a különböző, hang és videó alapú kommunikációs megoldások világunkban igen elterjedtek, az emberek közötti távoli kapcsolattartás többsége napjainkban is szöveges, írott formában történik. Az internet elterjedésével még a korábbanál is könnyebbé, sőt megszokottá vált az írott információ névtelen közlése, amelyen keresztül a személyazonosság akár még az igazságszolgáltatás számára sem, vagy

csak igen nehezen visszakövethető. Az anonim közlésen keresztül ugyan a személyes vélemény szabadabb módon adható át, hozzájárul a szólásszabadsághoz, de rengeteg veszélyt is hordoz. Ilyen esetben a szerzők kevésbé meggondoltan fogalmaznak, sőt akár mások jogait is megsérthetik. Szükség van tehát olyan megoldásokra, amelyekkel ezen visszaélések megnehezíthetők.

Egy írott szöveg szerzőjét felismerni pusztán a szövegre támaszkodva egyáltalán nem triviális feladat. A szöveg óhatatlanul tartalmaz azonban bizonyos nyelvi markereket, amelyek kifejezetten adott szerzőre jellemzőek; a cél ezeknek a nyelvi markereknek a kinyerése, összevetése és az eredmények kiértékelése – automatizált módon.

Két szöveg szerzőjének összerendelése - leegyszerűsítve két szövegtörzs összehasonlítása - továbbra sem egyszerű, a szövegben rengeteg különböző stilometriai jellemző feltárható, amelyek utalhatnak az egyezésre (Michell (2013); Coulthard (1994, 2004)). Ekkora mennyiségű adat pedig szabad szemmel nehezen nyerhető ki és rendszerezhető. Az automatizáció itt utat nyithat egy jelentősen objektívebb látásmód felé, amely kiküszöbölheti az emberi szemlélet és memória hiányosságait. Habár a kriminalisztikai nyelvi elemzés részleges automatizálása több országban már régóta jelen van, az ebben rejlő lehetőségek és lehetséges fejlesztések kiaknázása még messze el van maradva más nyomozati ágak modernizációs törekvéseitől. A német bűnügyi kutatóintézet (Bundeskriminalamt) adatain például történt már hasonló kísérlet (Ishihara (2017)), amelyben a szerző valószínűségi arányokkal (LR, likelihood ratio) végzett kísérletei alapján mutat rá egyes jellemzők információtartalmára a feladat szempontjából. Kísérleteik alapján például a szavankénti átlagos karakterszám, az írásjelek aránya és a szókincs nagysága jó mérőszámoknak bizonyulnak kisebb szövegminták esetén is.

Magyarországon a magyar nyelvű szövegek kriminalisztikai elemzése jelenleg kézzel történik, a szókészlet elemeinek feltárása és összehasonlítása pedig általános konkordancia programokkal (pl. Laurence Anthony szoftverei¹), ennek automatizációjára tudomásunk szerint korábban nem történt kutatás. A magyar nyelv természetesen továbbra is egyedi problémakört jelent, ami kihát a témában szokásos módszerek felhasználási lehetőségeire. Habár egyes jellemzők használata hasonlóan működik akár a használt nyelvtől függetlenül is, azok információtartalma mégis változhat. Hasonló kiértékelésre tehát szintén szükség van magyar nyelvre is. Ehhez először a nyelvi markerek megállapítására és automatizált kinyerésére kell helyeznünk a hangsúlyt. Az elemzés eredményei pedig a későbbiekben hozzájárulhatnak az intelligens megoldásokhoz is, amelynek során egy gépi megoldás képes lehet objektív módon segíteni nem csak a jellemzők átlátását, hanem magát az egyezésre vonatkozó döntést is.

2. Háttér

Habár külföldön széles körű tudományszervező és aktív kutatási tevékenység folyik a kriminállingvisztika területén (Coulthard és Johnson (2010); McMen-

¹ <https://www.laurenceanthony.net/software.html>

min (2002); Shuy (2006); Nini (2014); Olsson (2004)), a szövegek kriminalisztikai szempontú gépi elemzésének is nagy bibliográfiája van (pl. Crespo és Frías (2015); Ishihara (2010); Nirkhi és mtsai (2016); Rexha és mtsai (2018); Sousa-Silva (2018); Zhang és mtsai (2014)), sőt nyelveken átívelő megoldásokkal is kísérleteznek (Faqeeh és mtsai (2014); Llorens és Delany (2016)), a magyarországi igazságügyi nyelvész szakértők lényegében a kriminalisztikai szövegnyelvészet máig egyetlen összefoglaló kötete alapján végzik a tevékenységüket (Nagy (1980)). Maga a nyelvész szakértés mibenléte azóta is csak egy viszonylag szűk nyelvész/nyelvészeti érdeklődésű réteget, egy-két cikk, tanulmány erejéig foglalkoztat (pl. Szakácsné Farkas és Jánosné (1988); Pápay (2007); Szegedi (2018); Tolmainé Kabók (2015); Ürmösné Simon (2019); Ránki (2011)). Ennek oka, hogy Magyarországon összesen hat – az Igazságügyi Minisztérium szakértői nyilvántartásába – bejegyzett nyelvész szakértő tevékenykedik, forenzikus intézményben még kevesebb, nem erős a terület lefedettsége. Éppen ezért fontos az NBSZ-nek a szerzőségvizsgálat automatizációjának bevezetése, a megrendelők minél hatékonyabb kiszolgálása érdekében, és ebben megerősítenek minket a külföldi partnerszolgálatok tapasztalatai és a nemzetközi kutatási irányok is.

A kriminalisztikai vagy bűnügyi nyelvészet, még specifikusabban a forensic stylistic/forensic authorship research (Perlman (2018)) alapja az egyéni nyelvhasználat (idiolektus), mellyel nem csak a költők, írók rendelkeznek (Ürmösné Simon (2019)). Az idiolektus az, ahogyan a konkrét személy alkalmazza a nyelvet, mely magában hordozza a nyelvelsajátítás hogyanját és a személy nyelvhez való viszonyulását is. A szocializációs közegekben elsajátítottuk valamilyen mélységben a nyelvet, az adott nyelv (esetünkben a magyar) eszközkészlete valamilyen módon a rendelkezésünkre áll; a nyelv alkalmazása során a fejünkben lévő eszközöket variáljuk, válogatjuk, kombináljuk; a nyelvi kompetenciánkra hatással van szűk környezetünk, társadalmi pozícióink, tanulmányaink, olvasmányélményeink, fogékonyságunk a nyelvre, életkorunk, nemünk stb. (Szilák (1980)). Az egyén nyelvhasználatának bizonyos elemeit adott beszédhelyzethez, adott témához, adott műfajhoz, adott közönséghez stb. igazítja, ugyanakkor bizonyos elemei egyéni, mélyen rögzült, nem feltétlenül tudatos választás eredményei (jellemzően ilyenek a kapcsolóelemek, funkciószavak, mondatalkotási eljárás, állandósult szókapcsolatok használata, helyesírási esetek, ismétlődések, szókészlet egyes elemei stb.). A szakértői munka alapja a fogalmazó általános és különös stílussajátosságainak detektálása.

A nyelvész szakértő munkája során a fenyegetést, becsületsértést, zsarolást, rágalmozást megvalósító névtelen (kérdéses) írásművekből feltárható stílusjegyeket összeveti a gyanúsítottól származó szövegminták stílusjegyeivel, és a hasonlóságok/különbségek mennyiségi és minőségi mutatói alapján következtetéseket von le arra vonatkozóan, hogy a kérdéses írásművek fogalmazója a gyanúsított személy-e vagy sem. Az 1:1 alapú összehasonlítás a legtöbb kriminalisztikai szakértői gyakorlat alapja: adott kérdéses aláírást, okmányt, tárgyat, fotót, beszédet stb. adott mintához; esetünkben a kérdéses írásművek és az összehasonlító szövegminták összevetését jelenti elsősorban morfológiai, szintaktikai, szemantikai és pragmatikai szinten, kvalitatív és kvantitatív módon. A stilometrián felül az

általános korpusznyelvészeti tanulmányokhoz hasonlóan a nyelvész szakértő is alkalmazza azt a technikát, hogy a kérdéses írásművek egyes stílusjegyeit (pl. bizonyos szavak előfordulási gyakoriságát) hasonlítja egy általános korpuszban lévőkhöz (<http://mnsz.nytud.hu>), hogy megállapítsa, van-e jelentős eltérés a szokásos (átlagos) nyelvhasználattól az adott stílusjegy tekintetében vagy sem (lényegében a Magyar Nemzeti Szövegtárat referenciakorpusznak alkalmazva egy-egy szóelőfordulást, grammatikai jelenséget illetően).

Az online környezetben megvalósított bűncselekmények elkövetői informatikailag legtöbbször lenyomozhatók; ez esetben a nyelvész szakértő bevonása az inkriminált szövegek fogalmazójának megállapítása céljából nem feltétlenül szükséges, hiszen munkája időigényesebb, ezért drágább is. Azoknál a bűnelkövetőknél viszont, akik jól rejtik magukat a digitális térben, a nyelvész szakértők jelenthetik a megoldást, ugyanis képesek az egyre nagyobb számú internetes (verbális és írott formában elkövetett) bűncselekményekhez köthető, online, anonim vagy pszeudonim interakciók nyelvészeti elemzésére. Az internetes szövegek elemzésének igénye pedig új kihívások elé állította a szakértőket. A kommunikációs platformok megváltozásával az inkriminált szövegek is megváltoztak, így a hagyományos kriminalisztikai szövegnyelvészeti felfogás, elemzési módszer sok esetben nem alkalmazható (pl. egy ékezet nélkül írott szövegben az ékezethibák nem látszanak; az IM-szövegekben (Instant Messaging), kommentekben nincs relevanciája a különírás-egybeírásnak, a toldalékok elmaradnak, nincsenek írásjelek, gyakori az ismétlés és a rövidítés stb.), vagyis az idiolektus részének tekinthető nyelvi jellemzők egy része nem nyerhető ki a szövegből. A nyomozó hatóság kérdése a szakértőhöz azonban ezeknél a szövegeknél is változatlan: ki a névtelen szövegek írója?

Az NBSZ szakterület-fejlesztési irányai az alábbiakat tartalmazza:

1. Gépi szövegelemzés: Az első állomás az egy platformon, minél több nyelvi szinten megvalósuló gépi szövegfeldolgozás; automatizáció bevezetése azon szövegek esetében, melyek manuálisan, illetve a már meglévő szoftveres elemzőkkel feldolgozhatók, majd a tapasztalatok fényében megkísérelhető a rövidebb/hiányos szövegekből történő gépi adatkinyerés is.
2. Gépi szöveg-összehasonlítás: A gépi elemző által, mely több szövegszinten is megbízhatóan és hatékonyan dolgozza fel a szöveget, megvalósulhat egy automatikus 1:1 alapú szöveg-összehasonlítás humán, szakértői kontrollal: a gép által feltárt nyelvi jegyek kézzel történő priorálása, vagyis az adott nyelvi jegy megkülönböztető voltának, azonosító erejének meghatározása. A cél, hogy a tesztelést, finomítást követően ez a folyamat is teljesen automatikussá váljon.
3. Gépi szöveg-összehasonlítás meglévő adatbázison: A 2. pont megalapozhatja, előkészítheti az 1:N alapú összehasonlítást, melynek keretében az újonnan keletkezett kérdéses írásműveket össze tudjuk hasonlítani korábbi bűnügyekben keletkezett írásművekkel is, ismétlődéseket, sémákat keresve (pl. fenyegető levelek általános jellegzetességei), vagy akár elkövetői egybeeséseket felismerése is megtörténhet (többszörös bűnelkövetők: ugyanaz a személy

máskor, máshol, másokat, más okból fenyeget/zsarol/rágalmaz). A cél, hogy gépi úton hasonlósági sorrend felállítása történjen meg (score érték alapján).

4. LR-alapú értékelés: A bevezetőben említett, a Bundeskriminalamt által próbált LR-alapú 1:1 szövegösszehasonlítás más kriminalisztikai területen működik (Orbán (2018)), például a nyelvész szakértői területhez legközelebb álló hangtechnikai szakterületen is (pl. Fejes (2018)). A biometrikus azonosító rendszer működéséhez populációs adatbázis kiépítése szükséges, lévén a rendszer a kérdéses hanganyagot hasonlítja a hangmintához és a populációs adatbázisban lévő hanganyagokhoz, majd annak a valószínűségét adja meg, hogy melyiknek nagyobb az esélye: a kérdéses beszélő a mintaadó vagy inkább bárki más. A módszer tehát nagymértékben független a humán szakértőtől: két hipotézis közül a gép határozza meg a valószínűség mértékét, bűnügyi szempontból a gép dönti el, hogy a mintaadó személy bűnös (azonos a kérdéses beszélővel) vagy nem bűnös (a kérdéses beszélő a populációs adatbázishoz jobban illeszkedik). Ennek a módszernek a bevezetése a nyelvész szakértői területre még várat magára; a folyamatnál kulcsfontosságú a populációs adatbázis összetétele és a szövegekből az adatkinyerés hatékonysága, pontossága.

Jelen kutatásunk első állomása a gépi elemző és összehasonlító kidolgozásáról szól, a minél szélesebb körű és pontosabb adatkinyerés megvalósításáról, egyre kevesebb humán részvétellel, valamint arról, hogy a kinyert adatokat statisztikai szinten összehasonlíthatóvá tegyük a – pusztán statisztikai adatok alapján számított – gépi értéktételel megfogalmazásához a fogalmazó-szerzők azonosságának/különbözőségének valószínűségét illetően.

3. Módszerek

Jelen tanulmányban azt vizsgáljuk, hogy milyen típusú nyelvi markerek játszanak fő szerepet a szövegek szerzőinek azonosításában. Ehhez egy kisebb korpuszt készítettünk, mely négy különböző szerzőtől tartalmaz különböző írásműveket. Ezek között vannak “hétköznapi” jellegű dokumentumok is, de vannak különféle bűnügyekhez kapcsolódó írások is (pl. zsaroló vagy fenyegető levelek.) A korpusz alapvető adatai az 1. táblázatban láthatók.

1. táblázat. A kísérletek során felhasznált korpusz adatai.

Szerző	Tokenszám	Mondatszám	Dokumentumszám
A	7127	555	10
B	3205	170	6
C	6323	461	13
D	15791	912	32
Összesen	32446	2098	61

E tanulmányban arra keressük a választ, hogy akár egy kis elemszámú mintán is kimutathatók-e olyan (szignifikáns) különbségek, amelyek egyértelműen jellemzik egy szerző idiolektusát, ami a szerzőség megállapításában fontos szerepet játszhat. A szövegeket a magyarlanc nyelvi elemzővel (Zsibrita és mtsai (2013)) elemeztük morfológiai és szintaktikai szinten, valamint különféle szótárak alapján a szókincsüket is részletes elemzésnek vetettük alá. A kapott elemzésből automatikusan nyertük ki az alábbi jellemzőket.

Statisztikai jellemzők: tokenek száma, mondatok száma, lemmák száma és aránya, mondatok átlagos hossza, csupa nagybetűből álló szavak száma és aránya, nagy kezdőbetűs szavak száma és aránya, kijelentő mondatok száma és aránya, felszólító/felkiáltó/óhajtó mondatok száma aránya, kérdő mondatok száma és aránya, a szöveg telítettsége (lemmaszám / tagmondatok száma)

Morfológiai jellemzők: Főnevek, igék, melléknevek, ismeretlen szavak, határozószavak, tulajdonnevek, számnevek, névmások, vonatkozó és mutató névmások, névutók és központosítás száma és aránya, múlt és jelen idejű, feltételes és felszólító módú, gyakorító, műveltető és ható, adott számú és személyű igék száma és aránya, középfokú és felsőfokú melléknevek száma és aránya, többes számú főnevek száma és aránya, különleges képzővel/végződéssel rendelkező szavak száma és aránya

Szintaktikai jellemzők: Alanyok, tárgyak, jelzők, határozók, alárendelések, mellérendelések száma és aránya, tagmondatok száma és aránya, egyszerű mondatok száma és aránya, összetett mondatok száma és aránya, egy, két, három vagy négy tagmondatból álló mondatok száma és aránya

Szemantikai jellemzők: Pozitív és negatív töltetű szavak száma és aránya (Szabó (2015) alapján), negatív emotív szavak száma és aránya, tagadószavak, funkciószavak és tartalmas szavak száma és aránya, trágár és rasszista szavak száma és aránya, speciális stílusértékű szavak száma és aránya, megszólítások, elköszönő formulák és utóiratok száma és aránya, bizonytalanságra (Vincze (2014)) és érzelmekre (Szabó és Vincze (2016)) utaló szavak száma és aránya

Pragmatikai jellemzők: beszédaktusok száma és aránya, idézetek és gondolatjelek száma és aránya, kifelé és befelé forduló igék száma és aránya, meggyőzést jelentő igék száma és aránya, diskurzusjelölők száma és aránya

Ezekon felül még szókincselemzést is végeztünk a leggyakoribb szavak jelentésmezőinek összevetésével.

4. Eredmények

Az alábbiakban csak a legfontosabb eredményekre koncentrálnak mutatjuk be vizsgálataink eredményét. Mivel a szerzőktől eltérő nagyságú minta állt rendelkezésünkre, különös tekintettel arra, hogy szövegeink közel fele a D szerzőtől származik, elsősorban az egyes nyelvi jelenségek arányaira összpontosítunk, nem a darabszám szerinti előfordulásra. Ugyanakkor megemlítjük, hogy elvégzett statisztikai szignifikanciatesztjeink szerint a négy szerző nyelvhasználata szignifikáns eltérést mutat egymástól (ANOVA, $F=13,2948$, $p=2,23017E-08$). Az

alábbiakban felsorolt jellemzők kinyerése mind automatizált módon, szoftveresen történt, ezen jellemzők kinyerése tehát jól automatizálható.

4.1. Statisztikai jellemzők

A 2. táblázat mutatja a legfontosabb statisztikai jellemzők számszerűsített eredményeit. Érdekes megfigyelni, hogy itt is már nagy különbségek mutatkoznak a négy szerző között: B szerző kiemelkedően hosszú mondatokat ír, ami megmutatkozik az átlagos mondathosszban és a mondatonként tagmondatok számában is. A szerzőnél különösen nagy a felszólító mondatok száma, míg C és D szerző viszonylag gazdagabb szókinccsel rendelkezik a másik két szerzőnél.

2. táblázat. A jelentősebb statisztikai jellemzők eredményei.

Jellemző	A	B	C	D
Lemmák aránya	0,48	0,44	0,57	0,58
Mondathossz	11,49	19,33	12,79	16,89
Kijelentő mondatok aránya	0,56	0,35	0,41	0,69
Felszólító mondatok aránya	0,30	0,11	0,09	0,09
Kérdések aránya	0,07	0,10	0,05	0,01
Telítettség	2,58	4,06	3,47	3,29

4.2. Morfológiai jellemzők

A 3. táblázat mutatja a legfontosabb morfológiai jellemzők számszerűsített eredményeit. Az egyes szófajok használati aránya is változik szerzőnként, illetve szerzőpáronként: úgy fest, hogy B és C, valamint A és D szerzők szófajhasználatára viszonylag közel áll egymáshoz páronként. A múlt és jelen idő használata terén ugyanakkor B szerző tér el szignifikánsan a többiekétől: nála jóval gyakoribb a múlt idő, azaz talán gyakrabban emlegeti a múltbeli cselekvéseket, míg a többi szerző inkább a jelenre/jövőre fókuszál. Noha A szerzőnél figyelhattuk meg a felszólító/felkiáltó mondatok nagyobb gyakoriságát, a morfológiai elemzések szerint D szerző használ nagyobb arányban felszólító módú igéket. E két tényező összevetése mindenképpen további vizsgálatokat indokol. Az igei személyragozást vizsgálva is nagy különbségeket láthatunk a szerzők között. Míg B szerző szinte csak önmagáról, illetve harmadik személyekről beszél (a második személyű igealakok gyakorlatilag teljesen hiányoznak a szövegeiből), addig A és D szerző gyakran használ E/2. formulákat, azaz közvetben megszólítja a szövegek címzettjét. Nem zárhatjuk ki természetesen, hogy B szerző magázva szólítja meg a címzetteket, de ennek igazolása további vizsgálatokat kíván.

4.3. Szintaktikai jellemzők

A 4. táblázat mutatja a legfontosabb szintaktikai jellemzők számszerűsített eredményeit. Itt nagyobb különbségeket az alá- és mellérendelések, valamint a hatá-

3. táblázat. A jelentősebb morfológiai jellemzők eredményei.

Jellemző	A	B	C	D
Főnevek aránya	0,19	0,28	0,27	0,19
Igék aránya	0,15	0,07	0,09	0,14
Mellénevek aránya	0,07	0,11	0,10	0,07
Határozószavak aránya	0,13	0,05	0,06	0,11
Tulajdonnevek aránya	0,00	0,05	0,03	0,02
Kötőszavak aránya	0,08	0,05	0,04	0,07
Írásjelek aránya	0,19	0,25	0,27	0,19
Névmások aránya	0,08	0,04	0,04	0,09
Múlt idő aránya	0,24	0,40	0,23	0,22
Jelen idő aránya	0,67	0,49	0,66	0,71
Feltételes mód aránya	0,08	0,07	0,04	0,04
Felszólító mód aránya	0,07	0,04	0,08	0,12
E/1. aránya	0,28	0,33	0,25	0,26
E/2. aránya	0,14	0,00	0,05	0,18
E/3. aránya	0,40	0,39	0,44	0,35
T/1. aránya	0,05	0,00	0,09	0,03
T/2. aránya	0,00	0,00	0,00	0,00
T/3. aránya	0,05	0,16	0,07	0,10

rozók használata terén láthatunk. B és C szerző a többiekhez képest gyakrabban használ mellérendelést, míg A szerző inkább az alárendelést és a határozók használatát részesíti előnyben. Az egyes mondatok összetettségét leíró mutatók alapján elmondhatjuk, hogy A és D szerző inkább összetett mondatokat, míg B és C szerző inkább egyszerű mondatokat használ. Ugyanakkor B szerzőnél kiemelkedik a négy tagmondatos mondatok aránya, ami arra utal, hogy noha az egyszerűbb mondatok gyakoribbak nála, ha mégis több tagmondatos foglalt egybe, akkor az az átlagosnál hosszabb mondatot eredményez.

4. táblázat. A jelentősebb szintaktikai jellemzők eredményei.

Jellemző	A	B	C	D
Alárendelés aránya	0,05	0,02	0,03	0,04
Határozók aránya	0,05	0,01	0,02	0,03
Mellérendelés aránya	0,05	0,10	0,09	0,07
Tagmondatosok száma	2,15	2,10	2,08	3,02
Egyszerű mondatok aránya	0,47	0,61	0,55	0,44
Összetett mondatok aránya	0,53	0,39	0,45	0,56
Egy tagmondatos mondatok aránya	0,47	0,61	0,55	0,44
Két tagmondatos mondatok aránya	0,15	0,08	0,10	0,13
Három tagmondatos mondatok aránya	0,24	0,07	0,17	0,13
Négy tagmondatos mondatok aránya	0,09	0,15	0,08	0,12

4.4. Szemantikai jellemzők

Az 5. táblázat mutatja a legfontosabb szemantikai jellemzők számszerűsített eredményeit. Várakozásainkkal némileg ellentétben, alig találhatunk különbséget a szerzők között e téren, szinte csak minimális eltéréseket mutat egy-egy szemantikai jegy. Noha a kriminalisztikai jelleg miatt például várhatnánk a negatív érzelmekre (pl. düh, frusztráció, fenyegetés) utaló szavak felbukkanását a szövegekben, valójában ezek csak minimálisan fordulnak elő a szövegekben. Természetesen előfordulhat az is, hogy a rendelkezésre álló szentiment- és emóciósztárak nem ilyen típusú szövegekre lettek felkészítve, emiatt nem tudjuk azonosítani a szövegekben rejlő emóciókat, vagy pedig a minta kis mérete nem teszi lehetővé ilyen jellegű különbségek kimutatását. A szövegek részletesebb szemantikai vizsgálata tehát mindenképpen indokolt a jövőben.

5. táblázat. A jelentősebb szemantikai jellemzők eredményei.

Jellemző	A	B	C	D
Pozitív szavak aránya	0,03	0,01	0,03	0,03
Negatív szavak aránya	0,02	0,02	0,02	0,03
Tagadás aránya	0,03	0,01	0,01	0,02
Funkciósavak aránya	0,46	0,44	0,46	0,47
Tartalmas szavak aránya	0,54	0,56	0,54	0,53
Feltételes szavak aránya	0,01	0,00	0,00	0,01
Weasel szavak aránya	0,02	0,01	0,01	0,02
Peacock szavak aránya	0,01	0,00	0,00	0,00
Hedge szavak aránya	0,01	0,00	0,01	0,01
Doxasztikus szavak aránya	0,01	0,00	0,01	0,01
Szorongás szavainak aránya	0,01	0,00	0,00	0,00
Öröm szavainak aránya	0,01	0,00	0,01	0,01

4.5. Pragmatikai jellemzők

A 6. táblázat mutatja a legfontosabb pragmatikai jellemzők számszerűsített eredményeit. Míg a kifelé forduló igék aránya nem mutat lényegesebb eltérést, addig B szerző feltűnően kevés befelé forduló igét használ, ő tehát inkább a külvilágra, semmint önmagára fókuszál írásaiban. Ugyanakkor ő használja arányaiban a legtöbb meggyőzésre utaló szót, ami ugyanakkor azt erősíti, hogy saját véleménye mellett érvel, azt ebben a formában kifejezésre juttatva. Érdekes még megfigyelni a diskurzusjelölők eltérő gyakoriságát: A és D szerzők használják gyakrabban, míg B és C szerzőknél némileg háttérbe szorul a szerepük.

4.6. Szókészleti jellemzők

Az alábbi ábrákon bemutatjuk az egyes szerzők által használt szavak leggyakoribb elemeit, szófelhő formájában.



2. ábra: Bal oldal: C szerző szókincsének leggyakoribb elemei, Jobb oldal: D szerző szókincsének leggyakoribb elemei

hoz is hatékonyan hozzájárulhat. A jellemzők alapján az alábbiakat állapíthatjuk meg az egyes szerzőkre nézve. A szerző nyelvhasználatában a hétköznapi élet szókincsé dominál, valószínűleg kevésbé iskolázott a többi szerzőnél. Különösen nagy nála a felszólító vagy felkiáltó mondatok száma, gyakran használ E/2. ragozást, azaz direktben megszólítja az olvasót. Igeidőket tekintve inkább jelen vagy jövőbeli eseményekre fókuszál, kevésbé említi a múltat. Diskurszusjelölők is gyakran fordulnak elő a szövegeiben, gyakran használ összetett (alárendelő) mondatokat, így írásai inkább az előbeszédre emlékeztetnek, semmint hivatalos szövegekre. B szerző mondatai kiemelkedően hosszúak, gyakran használ mellérendelést, ugyanakkor inkább az egyszerűbb mondatstruktúrát részesíti előnyben. A többi szerzőhöz képest ő többször említi múltbeli eseményeket, szinte csak E/1. vagy E/3. ragozást használ, azaz nem közvetlenül egy másik (tegezett) személynek intézi a mondandóját. Tetten érhető nála a kívülágra való fókuszálás, ugyanakkor azon tendencia is, hogy szeretné meggyőzni a szövegek címzettjét a saját véleményéről, feltehetően érvelő jelleggel próbálja bizonyítani igazát. A szövegek tematikája esetében nagyon speciális, a vadászat szókincsének erős jelenléte mindenféleképpen különleges támpontot adhat a szerző kiletére irányuló nyomozás során. Ugyanakkor az is kiderül, a hivatalok és rendőrségi eljárások világa sem ismeretlen számára. C szerző - B-hez hasonlóan - szintén inkább egyszerűbb mondatokat, de több mellérendelést használ, szókincsé változatosnak mondható. Diskurszusjelölőket ő is ritkábban használ, nyelvezete hivatalosabbnak tűnik, mint pl. A szerzőé. Szókincsét tekintve nála is a munka és a bűnügyek világa dominál. Nyelvezete inkább B-éhez áll közelebb. D szerző szintén gazdag szókinccsel rendelkezik. Inkább összetett mondatokban kommunikál, relatíve sok diskurszusjelölőt alkalmaz. Ő is gyakran használ E/2. alakokat, azaz megszólítja a címzettet, a jelenre vagy jövőre fókuszál. E sajátságok inkább az előbeszédi kommunikációra emlékeztetnek, így D nyelvezete (és profilja) talán A-éhoz áll legközelebb. Szókincsének alapján ő is tisztában van a rendőrségi eljárásokkal.

Természetesen a fenti következtetések csak segítik a nyomozók munkáját, önmagukban nem rendelkeznek bizonyító erővel az eljárásokban. Munkacsopor-

tunk jelenleg azon dolgozik, hogy a fenti jellemzőket automatikusan kinyerve a szerzőktől rendelkezésre álló szövegekből, azokat számszerűen összehasonlítva és azt egy nyelvész szakértő elé tárva, a gépi úton kinyert adatok automatikus összevetésével segítsük a szerzőazonosítás folyamatát.

6. Összegzés

E munka célja annak megmutatása volt, hogy egy kis mintán is lehetséges automatikus nyelvi elemzési eszközökkel olyan összefüggéseket felállítani, amelyek segítik a kriminalisztikai célú szerzőazonosítást. Felhívtuk a figyelmet több olyan nyelvi jellemzőre, melyek már akár kis szövegminta alapján is árulkodónak bizonyulhatnak a szerző stílusára nézve, illetve segíthetik a más szerzőktől való elkülönítést. További célunk, hogy a szerzőazonosítás folyamatát minél inkább automatizáljuk, valamint meghatározzuk az elemzett adatok gépi összehasonlítása alapján a fogalmazó-szerzők azonosságának/különbözőségének valószínűségét, hozzájárulva ehhez a nyomozói munka hatékonyságához és sikerességéhez.

Köszönetnyilvánítás

A publikációban szereplő kutatást (amelyet a Nemzetbiztonsági Szakszolgálat és a Szegedi Tudományegyetem valósított meg) az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

Hivatkozások

- Coulthard, M.: On the use of corpora in the analysis of forensic texts. In: Forensic Linguistics: The International Journal of Speech, Language and the Law. vol. 1, p. 27–43 (1994)
- Coulthard, M.: Author identification, idiolect, and linguistic uniqueness. In: Applied Linguistics. vol. 25, p. 431–447 (2004)
- Coulthard, M., Johnson, A.: The routledge handbook of forensic linguistics (2010)
- Crespo, M., Frías, A.: Stylistic authorship comparison and attribution of spanish news forum messages based on the treetagger pos tagger. In: Procedia - Social and Behavioral Sciences. p. 198–204. No. 212 (2015)
- Faqeeh, M., Abdulla, N., Al-Ayyoub, M., Jararweh, Y., Quwaider, M.: Cross-lingual short-text document classification for facebook comments. In: International Conference on Future Internet of Things and Cloud. p. 67–98 (2014)
- Fejes, A.: Beszéd alapján történő személyazonosítás új kihívásai a kriminalisztikában. In: Magyar Rendészet. p. 117–126. No. 2 (2018)
- Ishihara, S.: E-mail authorship verification for forensic investigation. In: Proceedings of the 2010 ACM Symposium on Applied Computing (SAC). vol. 24. Sierre, Switzerland (2010)

- Ishihara, S.: Strength of forensic text comparison evidence from stylometric features: A multivariate likelihood ratio-based analysis. In: *International Journal of Speech Language and the Law*. vol. 24, p. 67–98 (2017)
- Llorens, M., Delany, S.: Deep level lexical features for cross-lingual authorship attribution. In: *Proceedings of the First Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine 2016)*. Padova, Italy (03 2016)
- McMenamin, G.: *Forensic linguistics. advances in forensic stylistics*. CRC Press (2002)
- Michell, C.S.: Investigating the use of forensic stylistic and stylometric techniques in the analyses of authorship on a publicly accessible social networking site (facebook). In: *Dissertation. University of South Africa* (2013)
- Nagy, F.: *Kriminalisztikai szövegnyelvészet*. Akadémiai Kiadó, Budapest (1980)
- Nini, A.: *Authorship profiling in a forensic context*. In: *PHD thesis* (2014)
- Nirkhi, S., Dharaskar, R., Thakare, V.: Authorship verification of online messages for forensic investigation. In: *Procedia Computer Science*. vol. 78, p. 640–645 (2016)
- Olsson, J.: *Forensic linguistics: An introduction to language, crime, and the law*. Bloomsbury Publishing, New York (2004)
- Orbán, J.: Bayes-hálók a bűnügyi tudományokban. In: *PhD értekezés*. Pécs (2018)
- Perlman, A.: What is forensic stylistics? (2018), <https://www.language-expert.net/category-stylistic-analysis>
- Pápay, K.: Valószínűségi skálák az igazságügyi nyelvészetben. In: *I. Alkalmazott Nyelvészeti Doktorandusz Konferencia kötet*. p. 102–113. MTA Nyelvtudományi Intézet, Budapest (2007)
- Rexha, A., Kröll, M., Ziak, H., Kern, R.: Authorship identification of documents with high content similarity. In: *Scientometrics*. vol. 115, p. 223–237 (2018)
- Ránki, S.: A kriminalisztikai szövegnyelvészet hazai kutatástörténete 1960-tól 1990-ig. In: *E-nyelvmagazin* (2011), <https://e-nyelvmagazin.hu/2011/08/31/a-kriminalisztikai-szovegnyelveszet-hazai-kutatastortenete-1960-tol-1990-ig/>
- Shuy, R.: *Linguistics in the courtroom: A practical guide*. Oxford University Press, New York (2006)
- Ürmösné Simon, G.: Miben segítik a nyelvi ujjnyomok a nyomozást? In: *Magyar Rendészet*. p. 65–75. No. 1 (2019)
- Sousa-Silva, R.: Computational forensic linguistics: An overview of computational applications in forensic contexts 5, 118–143 (12 2018)
- Szabó, M.K.: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái. nyelv, kultúra, társadalom. In: *Segédkönyvek a nyelvészet tanulmányozásához*. p. 278–285 (2015)
- Szabó, M.K., Vincze, V.: Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: *Távlatok a mai magyar alkalmazott nyelvészetben*. p. 282–292. Tinta pp., Budapest (2016)
- Szakácsné Farkas, J., Jánosné, V.: A kriminalisztikai nyelvész szakértő munkája. In: *Belügyi Szemle*. vol. 26, p. 93–98 (1988)

- Szegedi, Z.: Kriminalisztikai szövegnyelvészet. In: Doktori disszertáció. Budapest (2018)
- Szilák, J.: Az írásszokások néhány formai jegyének háttéréről. In: Belügyi Szemle. vol. 16, p. 67–68 (1980)
- Tolnainé Kabók, Z.: Interdiszciplináris kapcsolatok a rendészettudományok és az alkalmazott nyelvészet között – különös tekintettel a törvénytudományi nyelvészetre. In: Magyar Rendészet. p. 131–145. No. 5 (2015)
- Vincze, V.: Uncertainty detection in hungarian texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. p. 1844–1853. Dublin City University and Association for Computational Linguistics, Dublin, Ireland (2014)
- Zhang, C., Wu, X., Niu, Z., Ding, W.: Authorship identification from unstructured texts. In: Knowledge-Based Systems. vol. 66, p. 99–111 (2014)
- Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of hungarian. In: Proceedings of RANLP. p. 763–771 (2013)

SZINTAXIS, SZEMANTIKA

Jogi szövegek tezaurusz alapú osztályozása: egy nyelvfüggetlen modell létrehozásának problémái

Nyéki Bence

Nyelvtudományi Intézet
nyeki.bence@nytud.hu

Kivonat A cikkben jogi szövegek automatikus többcímkes osztályozását vizsgáljuk. A feladat nagy mennyiségű betanító adatot igényel, azonban ha az osztályozás kivitelezhető a többnyelvű EUROVOC tezaurusz terminusai alapján, akkor elméleti lehetőség nyílik arra, hogy egy meghatározott nyelvű korpuszon betanított osztályozó nyelvfüggetlenül működhessen. A bináris relevancia módszerén alapuló osztályozónkat horvát korpuszon tanítottuk be, és bár teljesítménye horvát szövegeken elfogadható, kis méretű annotált magyar mintánkra alkalmazva gyenge eredményt mutatott. Ennek legvalószínűbb oka a horvát és a magyar korpusz közötti különbség a terminus- és címkeeloszlás szempontjából.

Kulcsszavak: osztályozás, többcímkes, tezaurusz, EUROVOC, nyelvfüggetlen

1. Bevezetés

Az Európai Unió számára fontos feladatnak számít az egyes tagállamok nemzeti nyelvű jogi szövegeinek összegyűjtése és egységes feldolgozása. Az ennek megvalósítására irányuló munka eredménye például az EUR-Lex¹ weboldal, mely az Európai Unió 24 nyelvén írt hivatalos dokumentumokhoz biztosít hozzáférést, vagy az EUROVOC² többnyelvű tezaurusz. A MARCELL CEF Telecom³ projekt ugyancsak egy lényeges cél elérésére irányul: jogi szövegek automatikus fordításának fejlesztésére hét nyelv között (horvát, magyar, román, bolgár, szlovák, szlovén, lengyel). E projekt keretein belül már létrehoztak egységesen annotált korpuszokat az érintett nyelvek jogi szövegeiből (Váradi és mtsai, 2020). Bár a korpuszok részletes morfológiai és szintaktikai információt tartalmaznak, valamint a már említett EUROVOC tezaurusz és a IATE⁴ adatbázis terminusainak jelölését is kivitelezték, még nem minden érintett nyelv anyagainak annotációja teljes: a következő lépés a dokumentumok szövegszintű annotálása az EUROVOC tezaurusz 21 legfelső fogalmi kategóriájával (doménjével).

A szövegek automatikus címkézésének egyik módja egy osztályozó létrehozása gépi tanulás segítségével. Ehhez azonban már előre megjelölt betanító adatokra

¹ <https://eur-lex.europa.eu/homepage.html>

² <https://op.europa.eu/en/web/eu-vocabularies>

³ <https://marcell-project.eu/>

⁴ <https://iate.europa.eu/home>

van szükség. Ilyen adatokat jelenleg a horvát és a szlovén korpusz tartalmaz. A rendelkezésünkre bocsátott horvát korpusz szövegeinek nagyobb részéhez (pontosan 19.802 dokumentumhoz) már hozzárendeltek egy vagy több EUROVOC domént. Megkíséreltük felhasználni ezeket a horvát nyelvű anyagokat egy nyelvfüggetlen osztályozó betanítására, amely képes lenne a magyar (vagy elméletileg bármely más nyelvű) korpusz szövegszintű annotálására is. Ennek elméleti lehetőségét az EUROVOC terminusai teremtik meg: a tezaursz alapját a SKOS (*Simple Knowledge Organization System*) séma (Isaac és Summers, 2009) szerint fogalmak alkotják, amelyek egymással alá-fölé rendeltségi viszonyban vannak, továbbá konceptuális sémákba és doménekbe rendeződnek. Ezeknek a fogalmaknak a tezaursz által lefedett összes nyelven egy-egy (vagy akár több szinonim) terminus felel meg. Ez egyértelmű megfeleltethetőséget eredményez különböző nyelvek egyes kifejezései között. Amennyiben betanítunk egy tetszőleges nyelvű annotált korpuszon egy olyan osztályozót, amely csak a tezaursz fogalmainak nyelvfüggetlen azonosítóit veszi figyelembe, akkor az alkalmazhatóvá válik bármilyen más nyelvű szövegre is. Ennek feltétele persze, hogy az adott szövegben előzetesen azonosítókkel jelöljük meg az azon fogalmaknak megfelelő terminusokat, amelyeken az osztályozó betanult. Ez az annotáció rendelkezésre is áll a korpuszokban.

A cikk hátralevő része a következőképpen épül fel: a 2. részben röviden tárgyaljuk a kapcsolódó irodalmat, majd a 3. részben áttérünk a horvát és magyar korpusz jellemzésére. A 4. rész a horvát szövegeken betanított modelleknek a horvát validációs és tesztanyagokon, illetve egy kis méretű magyar mintán való kiértékelését közli. Az 5. rész összegzi az eredményeket.

2. Kapcsolódó irodalom

Mivel egy-egy dokumentumot az EUROVOC több felső fogalmi kategóriája is jellemezhet, a fent vázolt feladat többszintű és többcímű automatikus osztályozás kivitelezését teszi szükségessé. A gépi tanulásban erre két megoldási stratégiát különítenek el: problématranszformációt és algoritmusadaptációt (Tsoumakas és mtsai, 2010; Dharmadhikari és mtsai, 2011). Az előbbi lényege a feladat olyan lépésekre való lebontása, amelyekre alkalmazhatók egycímű algoritmusok, míg az utóbbi lényege az egycímű algoritmusok olyan átalakítása, hogy egy-egy osztályozandó objektumhoz több címkét is képesek legyenek társítani. A két stratégiához tartozó problémákról, módszerekről és értékelésükről ír (Tsoumakas és mtsai, 2010). A különböző módszerek hatékonyságát mérte (Dharmadhikari és mtsai, 2011). A kitűzött dokumentumklasszifikációs feladathoz választott módszer részletesebb ismertetésére a 4. pontban kerül sor.

Tezaursz alapú gépi tanulást alkalmaztak (Sabbagh és mtsai, 2018) vállalatok gyártási kapacitásának osztályozására. A kutatás anyagául különböző vállalatok weblapjaiból kinyert szövegek szolgáltak. A szerzők szerint a tezaursz alkalmazásának előnye abban áll, hogy jól szűrhetővé válnak a szövegek azon elemei, amelyek lényeges információt tartalmaznak az osztályozás szempontjából. Ezenkívül az osztályozó betanítása is a tezaursz segítségével történt (nem

előre annotált korpusz alapján): minden címkéhez félig automatizált módszerrel választottak ki és súlyoztak releváns fogalmi egységeket a tezauszából, konceptuális tereket hozva létre ezzel az egyes címkékhez. E munka keretein belül egy nyelvű tezausszal dolgoztak.

Ugyancsak meg kell említeni, hogy az elmúlt években több kísérlet történt az EUROVOC tezausz felhasználására jogi szövegek osztályozásának céljából (Steinberger és mtsai, 2012; Chalkidis és mtsai, 2019). E munkák azonban címkékként használták az EUROVOC fogalmait, nem pedig jellemzőkként. Következésképpen az így létrehozott osztályozók óriási (több mint 7.000 elemből álló) címkehalmazzal dolgoztak, míg a jelen cikkben közölt kutatásban mindössze a 21 felső fogalmi kategória alkotta ezt a halmazt. Továbbá a betanítás nem terminusokon alapult, hanem a szövegek szavain (*bag of words*) vagy azok beágyazásain. A JEX szoftver⁵ (Steinberger és mtsai, 2012) leginkább a kézi annotálás megkönnyítésére alkalmas eszköz, amely a bemeneti dokumentumok vektorait az egyes címkék profiljaival (azaz a tanító adatok alapján kiszámolt centroidokkal) veti össze, és az n legjobb címkét rendeli az adott dokumentumokhoz (n egy meghatározott szám, a szerzők által beállított alapértelmezett értéke a 6). Az algoritmus így minden dokumentumhoz rangsorolja a címkéket, de mindig n darabot ad ki, ami az esetek jelentős részében nem egyezik a helyes címkék számával. A címkék rangsorolását mások később mélytanulási technológiák segítségével kiviteleztek (Chalkidis és mtsai, 2019).

3. A korpuszok

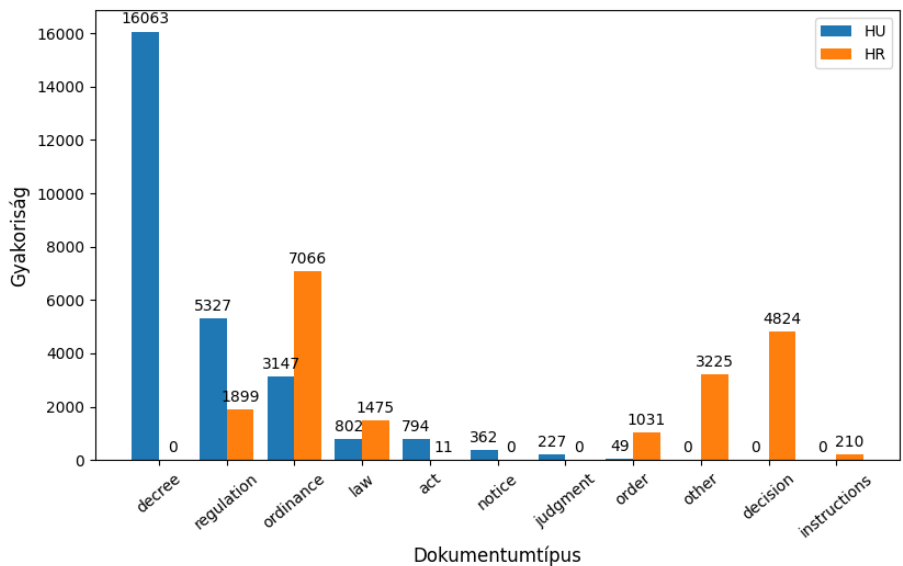
Ez a rész a horvát és a magyar jogi korpusz összevetését tartalmazza. A MARCELL projekt korpuszainak átfogó leírását adja (Váradis és mtsai, 2020).

3.1. Szavak és terminusok gyakoriságai

A magyar korpusz 26.821 dokumentumból áll, melyek döntő többsége határozat (16.063 szöveg tartozik ebbe a kategóriába). Ebből egyelőre mindössze 200 véletlenszerűen kiválasztott dokumentumot annotáltunk EUROVOC doménekkel. A teljes horvát anyag 33.559 dokumentumot foglal magába, a továbbiakban azonban csak annak a 19.802-nek adjuk jellemzését, amelyet kézzel megjelöltek az említett fogalmi kategóriákkal; ezekre fogunk horvát korpuszként hivatkozni, mivel a feladat szempontjából ezek relevánsak. Itt a leggyakoribb dokumentumtípus az utasítás, az idetartozó szövegek megközelítőleg harmadát teszik ki a korpusznak. A továbbiakban is láthatjuk, hogy a horvát minta több szempontból kiegyensúlyozottabb, mint a magyar.

Az 1. ábra a dokumentumtípusok eloszlását mutatja a magyar és horvát szövegekben. Mindkét korpusz dokumentumai tartalmazzák a megfelelő típus angol megnevezését, ezek a megnevezések láthatók az ábrán. Erre csak azok a típusok kerültek fel, amelyek legalább a két korpusz egyikében 100-nál többször fordulnak elő. A két korpusz típuscímkéi között nem teljes az átfedés.

⁵ <https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer>

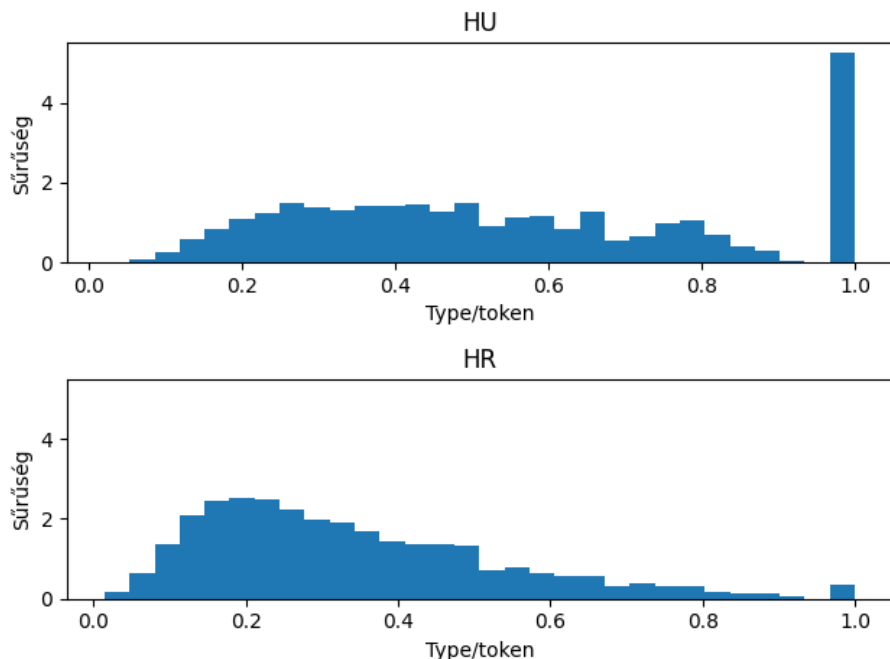


1. ábra: Dokumentumtípusok eloszlása

Az 1. táblázat néhány statisztikai adatot tüntet fel. Bár a horvát korpuszt mind az egyes dokumentumokban számolt szavak, mind az EUROVOC terminusok száma alapján nagyobb szórás jellemzi, az átlagok és mediánok is jóval magasabbak. Ezt a különbséget főként a magyar korpuszban nagy számban jelenlévő rövid (akár csak néhány mondatos) határozatok okozzák. Erre jobban rávilágít a 2. ábra, mely a két korpusz szövegeinek EUROVOC terminusai alapján kiszámolt type-token arányok hisztogramját mutatja (az egy dokumentumhoz tartozó type-token arányon a dokumentumban megtalálható különböző terminusok számának és a terminusok teljes számának hányada értendő). A hisztogram nem az egyes x-tengely menti szakaszokhoz tartozó abszolút gyakoriságokat, hanem az ezekből számolt sűrűséget tünteti fel. Ez azt jelenti, hogy az abszolút gyakoriságokat elosztjuk a gyakoriságok összegével és a hisztogram megfelelő oszlopainak szélességével, hogy az oszlopok területének összege 1 legyen.

1. táblázat. Statisztikai adatok gyakoriságokból

	HU		HR	
	szavak	terminusok	szavak	terminusok
Átlag	1164,23	64,62	2581	258,72
Medián	242	17	769	84
Szórás	3531,74	190,79	7411,67	635,55



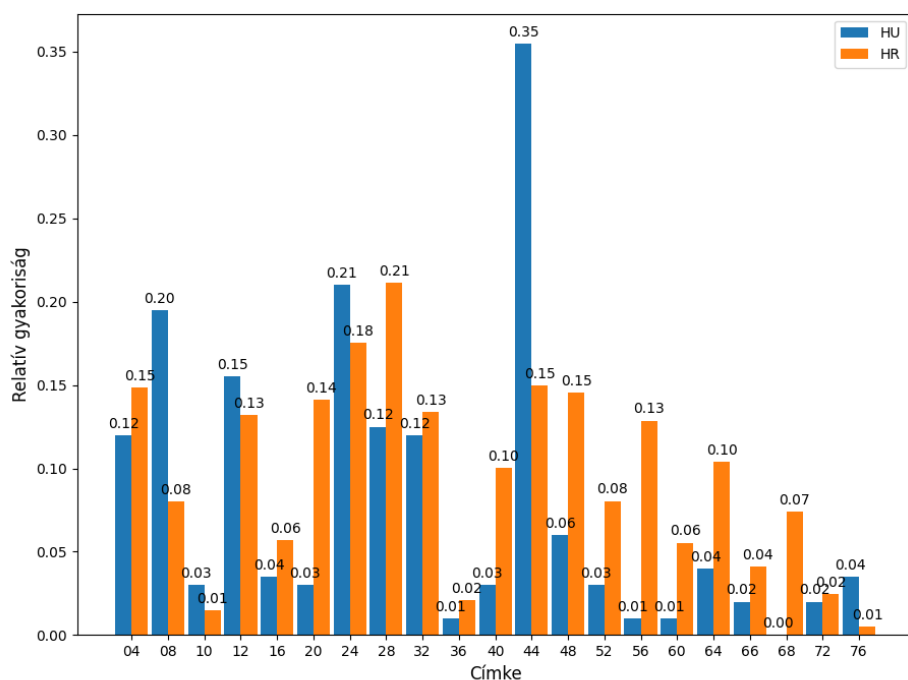
2. ábra: Type-token arányok

3.2. Címkék

Ugyancsak érdemes egy pillantást vetni az egyes címkék gyakoriságaira. A horvát korpuszból és a 200 kézzel megjelölt magyar szövegből kinyert relatív címkegyakoriságok⁶ a 3. ábrán láthatók. A relatív gyakoriságot úgy számolhatjuk, hogy elosztjuk az adott címke abszolút gyakoriságát a mintában lévő összes dokumentum számával, így jelen esetben a relatív gyakoriságok összege nagyobb, mint 1, hiszen egy dokumentumhoz több címke is tartozhat.

A horvát korpuszban a két leggyakoribb címke a 28 (társadalmi kérdések) és a 24 (pénzügyek): relatív gyakoriságuk 0,21, illetve 0,18 (mindegyik több mint

⁶ A feltüntetett kódok az EUROVOC doméneket jelzik. Feloldásuk a következő: 04: Politika, 08: Nemzetközi kapcsolatok, 10: Európai Unió, 12: Jog, 16: Közgazdaságtan, 20: Kereskedelem, 24: Pénzügyek, 28: Társadalmi kérdések, 32: Oktatás és kommunikáció, 36: Tudomány, 40: Vállalkozások és verseny, 44: Foglalkoztatás és munkakörülmények, 48: Közlekedés, 52: Környezet, 56: Mezőgazdaság, erdészet és halászat, 60: Agrárélelmiszer-ipar, 64: Termelés, technológia és kutatás, 66: Energia, 68: Ipar, 72: Földrajz, 76: Nemzetközi szervezetek. Lásd <https://op.europa.eu/en/web/eu-vocabularies/th-top-concept-scheme/-/resource/eurovoc/100141?target=Browse>.



3. ábra: Relatív címkegyakoriságok a korpuszokban

3.000 dokumentumban jelenik meg). Érdekes, hogy a 200 megjelölt dokumentumot tartalmazó magyar mintában a legtöbbször megfigyelhető címke gyakorisága sokkal inkább kiugró értéknek látszik: a 44-es EUROVOC domén (foglalkoztatás és munkakörülmények) relatív gyakorisága 0,35 (a címke 71-szer fordul elő), a második helyen álló 24-esé (pénzügyek) pedig 0,21 (ez 42-szeres előfordulást jelent). A legtöbb címke relatív gyakorisága a mintában nem éri el a 0,1-et, a 68-as (ipar) pedig teljesen hiányzik. Hangsúlyozni kell, hogy ezt a 200 dokumentumot (ami természetesen nem reprezentatív minta) egyetlen ember jelölte meg. Másik szakértő bizonyára részben más címkéket társított volna a dokumentumokhoz. Mindazonáltal ezek a számok tükrözik a magyar szövegek osztályozásának problémáját: a 44-es címke általában rövid kinevezéseket (pl. bírói tisztségre) jellemez. Ezekben a szövegekben ritkán fordul elő annyi terminus, hogy azok jól mutassák, melyik osztályba tartozik a dokumentum.

A type-token arányok és a címkegyakoriságok alapján arra következtethetünk, hogy a két korpusz összetétele erősen különbözik. Ezért a következőkben két problémát vizsgálunk:

1. Ha létrehozunk egy osztályozót, amely a horvát szövegeken tanult be, az mennyire hatékonyan fog további horvát szövegeket klasszifikálni?
2. Ugyanez az osztályozó használható lesz-e magyar szövegek feldolgozására is?

Egy horvát szövegeket hatékonyan osztályozó modell létrehozása tehát nem jelenti azt, hogy ugyanaz a modell a magyar szövegeket is jól fogja annotálni. Az alább bemutatott kísérletek éppen ezt támasztják alá.

4. Osztályozás

A következőkben az automatikus osztályozók kiértékelésére térünk át. Ehhez a horvát korpuszt dokumentumok véletlen kiválasztásával három részre osztottuk: tanító halmazra (15.842 dokumentum, kb. 80%), valamint validációs és teszhalmazra (mindegyik 1.980 dokumentumot tartalmaz, ez megközelítőleg 10-10%).

4.1. A címkék rangsorolása, egycímkés osztályozás

Ha ismert az egyes dokumentumokhoz társítandó címkék száma, a többcím-kés osztályozás megoldható a lehetséges címkék rangsorolásával aszerint, hogy mennyire jól illenek az adott osztályozandó dokumentumhoz. Az eddig tárgyalt feladat nem ehhez az esethez tartozik: egy szöveget a címkék tetszőleges nem üres részhalmaza jellemezhet, amelynek elemei nem rendezettek relevancia szerint. A címkék rangsorolása tehát nem tartozik a munka fő céljai közé. Mindazonáltal célszerű lehet néhány egyszerű, címkerangsorolásra irányuló kísérlet elvégzése. Ha ugyanis a nyelvfüggetlen többcím-kés osztályozást nem sikerül kielégítően kivitelezni, a rangsorolást viszont igen, a címkéket az egyes dokumentumokhoz relevancia szerint rendező algoritmus felhasználható egycímkés osztályozóként, ami voltaképpen félmegoldást jelentene. Ez nem valósult meg, de megfigyelhetjük, hogy a rangsoroláson alapuló egycímkés osztályozáskor és a többcím-kés klasszifikációkor ugyanaz a probléma merült fel: a rendelkezésre álló nyelvi anyagok különbözősége.

A címkék rangsorolásával kapcsolatos kísérlet lényege a következő: minden dokumentumhoz rangsoroljuk a címkéket, majd kiválasztjuk a legrelevánsabbat, és ellenőrizzük, az valóban eleme-e a valódi címkék halmazának (ha igen, az eredményt elfogadjuk).

Ennek megoldására két algoritmust dolgoztunk ki. Az egyik naiv, nem gépi tanulásra épülő módszer az EUROVOC tezaursz hierarchiáját igyekszik kihasználni. A SKOS struktúrában a fogalmak egyfelől hierarchikus viszonyban vannak egymáshoz képest (bővebb és szűkebb fogalmak szerint), másfelől elvontabb fogalmi sémák alá tartoznak. Utóbbiak az EUROVOC tezaurszban közvetlenül a legfelső kategóriák, a domének alatt helyezkednek el. Ha tehát minden dokumentumban megszámoljuk, hány terminus (vagyis ezeknek megfelelő fogalom) tartozik az egyes doménekhez, úgy azt a domént tekinthetjük a legjobb címkének, amelyet a legtöbb terminus reprezentál.

A másik lehetőségként egy naiv Bayes osztályozót vizsgáltunk, amely a nyers terminusgyakoriságokon tanult be (tehát egy olyan mátrixon, melynek sorai dokumentumoknak, oszlopai fogalmaknak felelnek meg, egyes elemei pedig egy adott fogalomnak megfelelő terminus dokumentumonkénti abszolút gyakoriságait mutatják). Az ehhez szükséges számítások elvégzéséhez úgy tekinthető, hogy

egy dokumentum beletartozik az összes olyan osztályba, amelynek címkéjével annotálták.

A hierarchia alapú osztályozó a teljes horvát korpuszon így 24,05%-os teljesítményt ért el, a 200 megjelölt magyar dokumentumon pedig 30%-ot. A feladat egyszerűségét figyelembe véve ezt az eredményt elégtelennek kell tekintenünk: nyilvánvaló, hogy egy gyakorlati alkalmazásban ennél jobb teljesítményt várnánk. A naiv Bayes elfogadhatóbb eredményeket adott a horvát tesztanyagokon (mivel itt csak egy modellel kísérletezünk, a validációs anyagok egyesíthetők voltak a tesztalmazzal): 77,88%-ot, a tanító halmazon pedig 79,72%-ot. A magyar mintán egészen más eredmény született: mindössze 24%. A naiv Bayes teljesítményére erősen negatív hatással lehet, ha a tanítóhalmazban az osztályok reprezentációja kiegyensúlyozatlan, ami magyarázhatja a horvát anyagon mért kb. 20% hibát. A magyar szövegeken kapott alacsony érték azt támasztja alá (bár nem bizonyítja), hogy a magyar minta a terminusok eloszlását tekintve más jellegű dokumentumokból áll, mint a horvát korpusz.

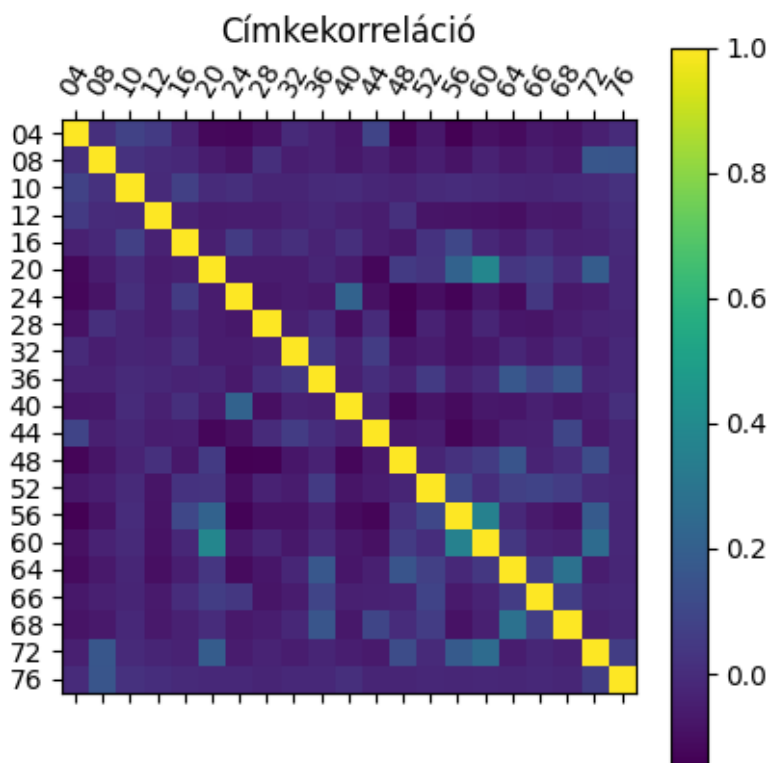
4.2. Többcímkes osztályozás

A relevánsabb feladat egy olyan osztályozó létrehozása volt, amely minden dokumentumhoz egy vagy több nem rangsorolt címkét társít. Ennek kivitelezésére a BR (*Binary Relevance*) vagy OvR (*One vs Rest*) módszert használtuk. Ez egy problématranszformációs módszer, melynek lényege, hogy minden egyes lehetséges címkehez felállít egy-egy bináris osztályozót, amely a többi címkétől függetlenül abban hoz döntést, hogy az adott címke jól jellemzi-e az osztályozandó dokumentumot vagy sem (Tsoumakas és mtsai, 2010; Dharmadhikari és mtsai, 2011). A bináris osztályozók bármilyen algoritmus alapján működhetnek. A természetes nyelvek feldolgozására gyakran alkalmazott naiv Bayes, SVM (*Support Vector Machine*) és k -legközelebbi szomszéd (KNN, *K Nearest Neighbors*) algoritmusokkal dolgoztunk. A modellek létrehozására a Python *scikit-learn* könyvtárát használtuk ⁷ (Pedregosa és mtsai, 2011). A BR előnye, hogy konceptuálisan egyszerű és nem igényel nagy számítástechnikai kapacitást, hátránya viszont az, hogy nem veszi figyelembe az egyes címkék közötti korrelációt. Ez komoly leegyszerűsítést jelenthet, bár ha vetünk egy pillantást a címkék korrelációs mátrixára (lásd 4. ábra), amelyet a teljes horvát korpusz alapján számoltunk ki, akkor csak viszonylag alacsony együtthatókat látunk. A legnagyobb korrelációs koefficiens a 60-as (agrárélelmiszer-ipar) és a 20-as (kereskedelem) címkék között számoltuk ki, de ennek értéke is kisebb, mint 0,4.

A választott módszer hátránya ellenére a horvát validációs és tesztanyagokon mért eredmények azt mutatják, hogy a BR alkalmazása elfogadható eredményekhez vezet.

Első lépésként a horvát szövegekből gyakorisági mátrixokat nyertünk ki, melyekben minden sor egy-egy dokumentumot, az oszlopok pedig egy-egy EURO-VOC fogalmat reprezentáltak: az egyes elemek így értelemszerűen azt mutatják, hányszor fordul elő valamely fogalom (azaz a neki megfelelő terminus) egy adott

⁷ <https://scikit-learn.org/stable/>



4. ábra: A horvát szövegek címkéi közötti korrelációs együtthatók

dokumentumban. Ezeket az adatokat előzetesen átalakítottuk. Ez az osztályozó létrehozásának rendkívül fontos lépése, mely erősen befolyásolhatja a végeredményt. A különböző paraméterű SVM modelleknek német nyelvű szövegeken való betanítása után (Leopold és Kindermann, 2002) például arra a következtetésre jutottak, hogy az előfeldolgozás (TF-IDF, lemmatizálás stb.) fontosabbak az eredmények szempontjából, mint a modell magfüggvényének választása.

Két előfeldolgozási módszert alkalmaztunk: TF-IDF-et L2 normalizálással és főkomponens-analízist. A TF-IDF lényege, hogy nem csak abszolút gyakoriságaik szerint súlyozza a terminusokat, hanem azt is számításba veszi, hogy hány különböző dokumentumban fordulnak elő: nagyobb súlyt kapnak azok, amelyek kevesebb dokumentumban jelennek meg. A főkomponens-analízis pedig olyan technika, mely a jellemzőteret kisebb dimenziójú térbe képzi le kismértékű in-

formációvesztéssel (magas megmagyarázott varianciával a jellemzőket tekintve). Az utóbbi eljárás hátránya, hogy az új, csökkentett dimenziójú mátrix néhány eleme 0-nál kisebb értéket fog felvenni, így a főkomponens-analízis naív Bayes osztályozóval nem használható.

Kétféleképpen értékeltük a modelleket. A (Tsoumakas és Mitsakis, 2010) által megadott képletek közül a klasszifikációs pontosságot (1. képlet) és az egyszerű pontosságot (*Accuracy*, 2. képlet) választottuk.⁸ Az előbbi nagyon szigorú, mivel csak azokat az eseteket fogadja el, amikor a modell által jósolt címkék halmaza pontosan megegyezik az adott dokumentumhoz tartozó helyes címkehalmazzal (ezért a továbbiakban az "egyezés" megnevezéssel fogunk rá utalni). Pontos egyezés gyakran két emberi szakértő annotációi között sem várható. Ezért a második érték jobban jellemzi a modelleket.

A megadott képletekben Z_i és Y_i az i -edik megfigyeléshez tartozó valós, illetve jósolt címkehalmazt jelenti, I pedig olyan függvény, melynek értéke 1, ha az argumentuma igaz, egyébként pedig 0.

$$\frac{1}{m} \sum_{i=1}^m I(Z_i = Y_i) \quad (1)$$

$$\frac{1}{m} \sum_{i=1}^m \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \quad (2)$$

A különböző előfeldolgozási eljárások után kapott legfontosabb eredményeket a 2. táblázatban közöljük. A feltüntetett SVM modellek magfüggvénye a radiális bázisfüggvény (*rbf*). Lineáris magfüggvénnyel is kísérleteztünk, de az gyengébb eredményeket hozott, amiket itt nem közlünk. Hasonlóképpen a k -legközelebbi szomszéd algoritmust magasabb k értékekkel is teszteltük (ezt az értéket a táblázatban a KNN rövidítés melletti szám jelzi), amelyek kevésbé voltak hatékonyak, mint az alább megadott KNN modellek. Ehhez az algoritmushoz a dokumentumok közelségét az euklideszi távolsággal mértük.

A következő táblázattal kapcsolatban meg kell még jegyeznünk, hogy azokban az esetekben, amikor főkomponens-analízist használtunk (a megfelelő oszlopot az angol *Principal Component Analysis* terminus PCA rövidítése jelzi), csak annyira redukáltuk a jellemzőteret, hogy az átalakítás után legalább 95%-os megmagyarázott varianciát kapjunk. Ehhez 956 dimenzióra volt szükség.⁹

A naív Bayes eredményei gyengék, de a másik két osztályozónak sikerült olyan paramétereket találni, amelyek mellett viszonylag jól működnek. Ami az előfeldolgozást illeti, a TF-IDF határozottan javította az eredményeket, a főkomponens-analízis viszont többnyire rontotta. A jellemzők súlyozása tehát kulcsfontosságú,

⁸ Ez nem összetévesztendő azzal a szintén pontossággént vagy precízióként ismert mértékkel, amely azt mutatja, hogy mekkora része helyes azoknak az értékeknek, amelyeket egy modell megjósol.

⁹ Az EUROVOC összesen 7214 fogalmat tartalmaz, ám ezek jelentős része nem szerepelt a betanító adatokban. Így főkomponens-analízis nélkül is mindössze 4.722 dimenziója volt a jellemzőtérnek.

2. táblázat. A BR egyes algoritmusainak eredményei a validációs halmazon

Algoritmus	TF-IDF	PCA	Egyezés	Pontosság
NB	+	-	0,2676	0,5561
NB	-	-	0,0561	0,3898
KNN 1	+	+	0,5727	0,7291
KNN 1	+	-	0,5803	0,7317
KNN 1	-	+	0,4581	0,6008
KNN 1	-	-	0,4753	0,6536
KNN 5	+	+	0,4753	0,6536
KNN 5	+	-	0,4672	0,6467
KNN 5	-	+	0,3732	0,5141
KNN 5	-	-	0,4040	0,5521
SVM	+	+	0,5096	0,6880
SVM	+	-	0,5051	0,6844
SVM	-	+	0,1000	0,1623
SVM	-	-	0,1384	0,2148

dimenziócsökkentésre azonban nem volt szükség. Érdekes, hogy a legjobb teljesítményt a legegyszerűbb modell érte el: a k -legközelebbi szomszéd $k = 1$ paraméterrel. Ez azt jelenti, hogy a modell minden új dokumentumhoz megkeresi a jellemzőtérben a hozzá legközelebbi dokumentumot a tanító halmazból, és annak címkéit rendeli hozzá. Emellett az SVM sem teljesített rosszul a TF-IDF-fel és főkomponens-analízissel feldolgozott dokumentumokon (50% fölötti egyezés semmiképpen sem tekinthető gyengének). A tanító halmazon történő kiértékeléskor ugyanezek a modellek bizonyultak a legsikeresebbeknek: magától értetődő, hogy ebben a halmazban minden dokumentum legközelebbi szomszédja önmaga, ezért a KNN sosem hibázik. A jellemzők TF-IDF szerinti súlyozása és dimenziócsökkentés után az SVM által elért egyezés értéke 0,6458 volt, a pontosságé pedig 0,8065.

A táblázatban kiemelt két modellt a horvát tesztanyagokon és a magyar mintán is kipróbáltuk. Az így kapott értékeket a 3. táblázatban közöljük.

3. táblázat. A két legjobb modell kiértékelése a tesztalmazokon

Modell	HR		HU	
	Egyezés	Pontosság	Egyezés	Pontosság
KNN	0,5823	0,7372	0,0500	0,1628
SVM	0,5010	0,6894	0,0600	0,1446

A horvát tesztalmazon majdnem ugyanazokat az eredményeket kaptuk, mint a validáción, a magyar mintán azonban mindkét modell pontatlan, akárcsak a címkék rangsorolása esetében. Érdekes, hogy az SVM a magyar dokumentumok feléhez egyáltalán nem tudott címkét társítani: ez azt jelenti, hogy a BR összes

bináris osztályozója negatív döntést hozott, azaz irrelevánsnak értékelte azt a címkét, amelyikre betanult. Az SVM legtöbbször a 24-es (pénzügyek) címkét osztotta ki: 30-szor, míg a kézi annotációban ez többször (42-szer) szerepel. A KNN minden dokumentumhoz társított legalább egy címkét, de az eredményekből látható, hogy ezzel nem ért el jelentősen jobb teljesítményt. A KNN által a magyar szövegekhez leggyakrabban (62-szer) hozzárendelt címke a 04-es (politika), míg a kézi annotációban ez jóval ritkábban, 24-szer figyelhető meg. Az emberi jelölés szerint legtöbbször előforduló 44-es (foglalkoztatás és munkakörülmények) csak 20-szor jelent meg a KNN által jósolt címkék között.

Ugyancsak érdemes megfigyelni, hogyan változnak az eredmények, ha a magyar mintából csak a legtöbb EUROVOC terminust tartalmazó dokumentumokat vizsgáljuk. A csökkenő terminusgyakoriság szerint sorba rendezett magyar annotált szövegek közül az első 50 mindegyike legalább 49 terminust tartalmaz, a terminusgyakoriságok átlaga pedig 168,14 erre az 50 dokumentumra számolva. Csak ezeket figyelembe véve valamivel jobb eredményeket kapunk: a KNN 0,1 értékű egyezést és 0,2457-es pontosságot ad, ugyanezen mutatók értéke az SVM kimenetén pedig 0,14 és 0,2217. Bár ezek sem nagy számok, azt mutatják, hogy a rövidebb, kevés terminust tartalmazó magyar dokumentumokat (ahogy ez várható is) nehezebben kezelik a modellek.

Eddigi kísérleteink leírásának ezzel végére értünk. A következő feladatok között szerepel egy nagyobb magyar minta kézi annotálása, hogy így lehető legyen pontosabb képet kapni a létrehozott osztályozók működéséről magyar szövegeken.

5. Összegzés

Sikerült tehát egy egyszerű és megbízható (bár még fejleszthető) modellt létrehozni horvát jogi szövegek osztályozására. Az annotált magyar mintánkon azonban nem sikerült kielégítő eredményt elérni. A minta kis mérete miatt ezt nem tudjuk pontosan megmagyarázni, mindazonáltal a két korpusznak a 3. pontban leírt összevetését is figyelembe véve nagyon valószínű, hogy a dokumentumgyűjtemények közötti különbségek akadályai lehetnek egy nyelvfüggetlen osztályozó megalkotásának. Amennyiben a már rendelkezésre álló korpuszok valóban jól tükrözik a terminus- és címkeeloszlásokat az egyes nyelvekben, úgy a horvát jogi szövegeken aligha tanítható be olyan tezausz alapú osztályozó, amely a magyar szövegeket is helyesen kezeli. Mindazonáltal ahhoz, hogy általánosan értékelhessük a tezausz alapú nyelvfüggetlen osztályozókat, a megkezdett munka folytatásaként több különböző nyelv korpuszára és több nyelvpár anyagain végzett kísérletekre lesz szükség.

Köszönetnyilvánítás

A szerző köszönetét fejezi ki Marko Tadićnak és Vanja Štefanecnek a horvát korpusz megosztásáért, valamint Sass Bálintnak és Héja Enikőnek szakmai támogatásukért, javaslataikért.

Hivatkozások

- Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Androutsopoulos, I.: Large-Scale Multi-Label Text Classification on EU Legislation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6314–6322. Association for Computational Linguistics, Florence, Italy (07 2019), <https://www.aclweb.org/anthology/P19-1636>
- Dharmadhikari, C., Ingle, M., Kulkarni, P., Dharmadhikari, S.: A Comparative Analysis of Supervised Multi-label Text Classification Methods. *International Journal of Engineering Research and Applications* 1, 1952–1961 (2011)
- Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C (2009)
- Leopold, E., Kindermann, J.: Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning* 46, 423–444 (01 2002)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830 (2011)
- Sabbagh, R., Ameri, F., Yoder, R.: Thesaurus-Guided Text Analytics Technique for Capability-Based Classification of Manufacturing Suppliers. *Journal of Computing and Information Science in Engineering* 18 (03 2018)
- Steinberger, R., Ebrahim, M., Turchi, M.: JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool. In: LREC’2012 (2012)
- Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining Multi-label Data. In: Maimon, O., Rokach, L. (szerk.) *Data Mining and Knowledge Discovery Handbook*, pp. 667–685. Springer US, Boston, MA (2010)
- Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Niton, B., Ogródniczuk, M., Pkezik, P., Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Pais, V.F., Tufis, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J.: The MARCELL Legislative Corpus. In: LREC (2020)

Egy nagyobb magyar UD korpusz felé

Novák Attila^{1,2}, Novák Borbála^{1,2}

¹Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar

²MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport

Budapest, Práter u. 50/a.

{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat A cikkben egy olyan munkálat eredményeiről számolunk be, amelynek keretében a teljes Szeged Dependencia Treebanket megpróbáljuk a Universal Dependencies projekt keretében megfogalmazott annotációs elveknek megfelelő annotációjú korpuszá átalakítani, miközben az eredeti korpuszban szereplő annotációs hibákat, illetve következetlenségeket is igyekszünk feltárni, kijavítani, illetve megszüntetni.

Kulcsszavak: függőségi annotáció, Szeged Dependencia Treebank, Universal Dependencies,

1. Bevezetés

Két évvel ezelőtt egy olyan munkálatról számoltunk be (Novák és mtsai, 2019), amelynek keretében egy olyan függőségi alapú annotációs séma kialakítására tettünk kísérletet, amely a magyarra (illetve általában) korábban használt sémáknál jóval részletgazdagabb elemzést tartalmaz. Konkrétan az volt a célkitűzésünk, hogy az annotáció alkalmas legyen arra, hogy releváns kérdéseket lehessen a felhasználásával az adott szöveggel kapcsolatban feltenni. Így az annotációban használandó megkülönböztetések létjogosultságát is alapvetően az határozza meg, hogy az adott konstrukcióval kapcsolatban milyen kérdéseket lehet feltenni.

Akkor kiindulási anyagként a Universal Dependencies (UD) korpusz (Nivre és mtsai, 2020) 1800 mondatból (42000 token) álló magyar alkorpuszát használtuk (Vincze és mtsai, 2017), hogy nemzetközi szinten elfogadott annotációs sémából induljunk ki. Az UD korpusz nagyjából egységes elvek és kategóriák felhasználásával sok nyelv szövegeire tartalmaz morfoszintaktikai és szintaktikai függőségi elemzést. Mivel a magyar alkorpuszban szereplő annotáció sok szempontból nem felelt meg az érvényes UD specifikációnak, illetve sok véletlenszerű annotációs hibát találtunk, ezért már akkor foglalkoztunk a hibajavítással. A szerkezetek egy részét (pl. **appozitív szerkezetek** (*Katona Kálmán közlekedési minisztert*)), **egyeztetett predeterminánst** tartalmazó szerkezetek (*azt a kutyát*), **birtokos szerkezetek**, **névutós szerkezetek**) programozottan alakítottuk át. Más esetekben (pl. a harmadik személyű **névszói állítmányt** tartalmazó tagmondatok annotációjában az alany és az állítmány sok esetben meg volt cserélve) félig manuális módszerrel tudtuk javítani az annotációt: kézzel jeleltük meg a hibás mondatokat, ahol aztán az alany és állítmány annotációját programozottan javítottuk.

Miután a magyar UD alkorpusz annotációját kibővítettük a Novák és mtsai (2019)-ben említett annotációs elemekkel (a vonzatok és a szabad határozók egy meghatározott körének tematikus szerepére vonatkozó annotációval) kiegészítettük, megnéztük, hogy ezen az annotáción egy korszerű neurális függőségi parsert betanítva milyen teljesítményt tudunk elérni (Dozat és mtsai (2017)). Az eredmény kiábrándító volt (LAS=0.57). Ugyanakkor a korpusz mérete (a tanítóanyag mindössze 900 mondat) az UD specifikációban definiált alaprelációk tekintetében (ha olyan megkülönböztetéseket nem vesszük figyelembe, mint pl. a vonatkozó mellékmondatok megkülönböztetése `acl:relcl` az `acl` reláción belül) sem tesz lehetővé LAS=0.8 fölötti eredményt a legkorszerűbb parserek esetében sem, illetve a csak a tartalmas szavakat és a morfológiai annotációt is figyelembe vevő MLAS pontszám esetében MLAS=0.67 a legjobb eredmény (Zeman és mtsai, 2018). Ez az eredmény indított arra, hogy belekezdjük a teljes Szeged Dependencia Treebank Vincze és mtsai (2010) UD-vel kompatibilis formára hozásába.

2. Problémák és megoldások

A teljes Szeged Dependencia Treebank (SZDT) mintegy 82000 mondatból (1,5 millió tokenből) áll, így ígéretes méretű anyagnak tűnik egy viszonylag pontos függőségi elemző betanításához. Ugyanakkor számos az eredeti korpuszal, illetve az abban szereplő annotációval kapcsolatos probléma nehezíti a megfelelő minőségű UD-kompatibilis korpuszváltozat létrehozását. Alább áttekintjük ezeket a problémákat, illetve hogy milyen megoldást próbáltunk alkalmazni rájuk.

2.1. Hibás szóalakok

Nincs elemzés A helyesírási hibát tartalmazó szövegrészek nincsenek morfológiailag annotálva – legalábbis ahol a hibát az annotációs folyamat során észlelték. Ezekben az esetekben az morfológiai annotáció mindössze annyit tartalmaz, hogy a szó ‘hibás’ ((1) 2. token) vagy ‘az adott kontextusban hibás’ ((1) 3-4. token), de se elemezve, se lemmatizálva nincs.

```
(1) # sent_id = 10elb.ud - 11
    # text = De végülis oda értünk, mert jött az egyik osztálytársam ...
    1 De de CONJ CONJ _ 4 CONJ _ _
    2 végülis végülis X X _ 4 MODE _ _
    3 oda oda X X _ 4 PREVERB _ _
    4 értünk értünk X X _ 0 ROOT _ _
    ...
```

Sok esetben a hiba nem is az eredeti szövegből származik, hanem a feldolgozási folyamatba csúszott hiba eredménye (pl. az 1984 alkorpuszban minden kurzív szövegrész kezdetén álló szó a feldolgozás során egybeíródott az előző szóval, és ennek a szisztematikus alkorpusz-specifikus tokenizálási hibának a következményeit később nem javították ((2) 2. és 10. token).

- (2) # sent_id = 1984.ud - 683
 # text = Winston arégi számok-at tárcsázta a telekén, és kérte aTimes meg-
 jelölt számait.
 1 Winston Winston PROP_N PROP_N Case=Nom|Number=Sing 4 SUBJ _ _
 2 arégi arégi X X _ 3 ATT _ _
 3 számok-at számok-at X X _ 4 OBJ _ _
 ...
 10 aTimes aTimes X X _ 12 ATT _ _
 ...

A *Piszkos Fred* alkorpusz esetében „az író szándéka” vezetett rengeteg agrammatikus szöveg létrehozásához, ez az anyag azonban nem különösebben hasonlít ahhoz, ahogy valódi emberek hibáznak, tehát az sem világos, hogy a parser mit is tanulhatna pontosan ezekből a szövegrészekből.

A probléma nagyságrendjét érzékelteti, hogy a korpusz mondatainak több mint 10%-a tartalmaz hibásként annotált, morfológiailag nem elemzett szóalakokat.

Zárójelbe tett hibák Az iskolások fogalmazásainál az adatbevitel során is torzult az anyag (azon túl, amikor a diákoknak nem sikerült azt leírniuk, amit akartak). Egyrészt bizonyos szövegrészeket nem sikerült az adatbevivőknek elolvasni, és itt néha értelmezhetetlenül hiányos vagy torz mondatok kerültek a korpuszba (3a). A tanulók által zárójelbe tett szövegrészek ezzel szemben mind bekerültek a korpuszba, bár a diákok a zárójelbe tétellel a legtöbb esetben azt jelölték, hogy a szövegrész törlendő (3b). Később az annotátorok kénytelenek voltak ezekkel a részekkel is kezdeni valamit. Az SZDT függőségireláció-készletében nincs olyan elem, ami kifejezetten a hibás elemek megjelölésére szolgál. A zárójelbe került kifejezés feje ‘hibás’ „szófajcímke” kapott, és az APPEND relációval csatolták a mondat többi részéhez (pl. a (3b) mondatban a *nemcsa*, az *Én* ennek az alanyaként van megjelölve). Az APPEND reláció ugyanakkor a legtöbb esetben valóban a szöveg részét képező elemeket jelöl, sokszor még a ‘hibás’ „szófajcímke” és APPEND relációval csatolt szavak esetében is. A ‘hibás’ „szófaj” APPEND relációval csatolt fejlő zárójelbe tett szövegrészeket azonban nagy biztonsággal törölni lehet a szövegből. Legjobban ebben az esetben járunk, mert a ‘hiba’ mint szófajcímke megtartásának nincs értelme, és az iskolai kontextuson kívül nem fordul elő, hogy a zárójelbe tett részek azt jelentenék, hogy úgy kell érteni, mintha az a szövegrész oda se lenne írva, ezért nem igazán van értelme erre betanítani egy parsert.

- (3) a. (10erv.ud - 3158) Továbbá még véleményezem azt is, hogy a reális **XXX** sem **XXX** tárgyakkól kevesebb ill. a humán **XXX** fordítva.
 b. (10elb.ud - 7819) (**Én nemcsa**) Egy vasárnapi nap volt.

Nem annotált hibák Az elírások egy részét nem vették észre az annotáció során. Ezekben az esetekben a tokenizálás, illetve a lemma hibás.

Szétvágott és egyben maradt mondatok A korpusz elírásokból és az automatikus mondatrabontás hibáiból fakadóan 95 olyan mondatnak annotált egységet tartalmazott, amely valójában több mondatból állt. Emellett néhány olyan esetet is találtunk, ahol egy összefüggő mondat szakadt meg egy szó közepén.

Megoldás A hibás szóalakokkal kapcsolatos problémák megoldására azt láttuk célravezetőnek, hogy létrehozuk a hibás szövegek javított változatát, és ezt annotáljuk (nyilvántartva, hogy mit javítottunk). A hibásnak annotált szóalakot tartalmazó mondatokat kigyűjtöttük a forrásfájl és a mondatazonosító megjelölésével, és létrehoztuk az eredeti mondat normalizált változatát olyan formában, hogy abból rekonstruálható, hogy az eredeti mondatban mely tokeneket módosítottuk, töröltük, vontuk össze, választottuk részekre, illetve hova szúrtunk be esetleg új tokenet. A javított/normalizált változat kezdeti verzióját a gyakori hibákhoz készített automatikus hibajavítási lista alapján generáltuk az eredetiből (tehát ez már tartalmazott lényegében biztosra vehető javításokat). A hibásnak jelölt szóalakok ki voltak emelve, de a mondatokban észrevett egyéb hibákat is korrektúráztuk. Ebben a fázisban a mondathatárok javításán túl 9400 javítást vezettünk be az anyagba (1367 betűköztörlés, 1051 szó/tokentörlés, 1382 betűközbeszúrás, 321 szóbeszúrás, 5281 szójavítás).

A javítási lista alapján programozottan vezettük vissza a korpuszba a javításokat. A szétvágott elemeket jobb fejűnek feltételeztük és az elemek között alapesetben NE, illetve ATT relációt feltételeztünk attól függően, hogy névnek vagy más elemnek tűnt az eredeti token. A módosított tokeneket morfológiaiilag elemeztük, és a korpusz többi részén betanított PurePos taggerrel (Orosz és Novák, 2013) egyértelműsítettük. Az automatikus morfológiai annotáció és a függőségi címkék kézi átnézése/javítása folyamatban van.

2.2. Az igék, igenevek és ragozott névmások annotációja

Igenevek Az igenevek minden esetben melléknévként vagy határozószóként vannak annotálva. Az eredeti SZDT-ben az annotációból az sem derül ki, hogy igenévről van szó. A magyar UD korpuszban az igenév típusára utaló jegy szerepel morfoszintaktikai jegyek között. Ez sem volt azonban elegendő, amikor a bevezetésben említett kutatásunkban (Novák és mtsai, 2019) igei vonzatkereteket próbáltunk a szövegre illeszteni. Az illesztéshez az igenév igei tövére is szükség van az igenév típusa mellett a megfelelő vonzatkeret azonosításához. Természetesen problémát jelent, hogy számtalan lexikalizált melléknév van, amelyek formailag azonosak valamely igenévvvel, azonban ezek az igenevektől eltérően általában predikatívan is használhatóak (pl. *derült*, *tartózkodó* stb.) Ezen elemek nem predikatív előfordulásainak egyértelműsítése azonban sajnos csak manuális ellenőrzéssel valósítható meg.

„Ható”, műveltető és gyakorító igealakok Bizonyos teljesen produktív igei végződések (az egyértelműen inflexiós jellegű *-hat* végződés mellett pl. a műveltető- és a gyakorítóképző) nincsenek leválasztva a lemmáról, és nem jelennek meg a morfológiai elemzés szintjén (4).

- (4) adhat adhat VERB VERB Definite=Ind|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act

Ez az igenevek elemzetlenségéhez hasonlóan szintén problémát jelent a vonzatkeret-illesztés szempontjából. A *-hAt* végződés le nem vágása a lényegében önállósult lexikai elemmé vált *lehet* igétől eltekintve nem igazán tűnik indokoltnak. A gyakorító-, illetve műveltetőképzős alakok között ennél jóval több lexikalizált elem található (*mosogat*, *beírat stb.*), tehát ezek egyben tartása az eredeti SZDT-ben motiváltabb döntés volt.

Ragozott névmások A ragozott névmások annotációja nem felel meg sem a magyar morfológiai elemzők által adott, sem az UD specifikációban szereplő névmás-annotációs elveknek sem a lemma, sem a morfoszintaktikai jegyek tekintetében (pl. nem derül ki az eset).

- (5) nekem neki ADV ADV Number=Sing|Person=1|PronType=PrsPron

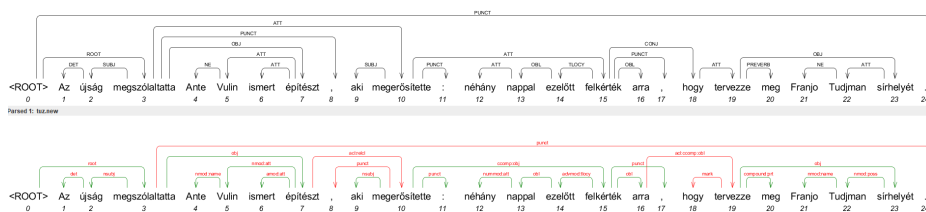
Egyéb szófajproblémák Sok funkciószó szófaji annotációja nem felel meg az intuíciónknak, illetve a morfológiai elemző által adott elemzéseknek (főleg a névmási jellegű és a kötőszószzerű elemek, illetve egyes névszói eredetű névutószzerű vonzatos határozószók esetében).

Megoldás A korpusz anyagát morfológiaileg újraelemeztük, és a meglévő annotációban az egyes tokenekhez rendelt lemma és morfoszintaktikai annotáció alapján kiszámítható annotációhoz leghasonlóbb, illetve azzal kompatibilis (ideális esetben azzal azonos) elemzést választottuk. Sajnos vannak olyan lényeges többértelműségek, amelyeket az eredeti annotáció neutralizál (pl. melléknévként lexikalizálódott igenevek, névutószzerű vonzatos határozószóként vagy időhatározóként lexikalizálódott ragozott főnevek stb.) Sajnos a szótípusok szintjén is tízezres nagyságrendű listát kell átnézni, ez jelenleg folyamatban van. (Annak idején az e-magyar projekt (Várad és mtsai, 2017) keretében készült egy konverzió az SZDT anyagáról az emMorph (Novák és mtsai, 2017) elemző által használt formátumra, azonban abban sok a hiba, sok esetben nem is a korpuszban szereplőnek megfelelő elemzésre történt a leképezés, ezért nem ebből az anyagból indultunk ki). A nem egyező elemzések fele a névelemek annotációjához köthető (l. a 2.4. részben).

2.3. Függőségi relációk

Egy-több megfeleltetés a relációk között A UD-ben és a SZDT-ben használt függőségi relációk halmaza nem feleltethető meg egymásnak egyértelműen. Az alap UD készlet pl. számos tagmondatok közötti relációs viszonyt megkülönböztet (az alárendelt tagmondat mondatbeli szerepétől függően), ugyanígy a frázisszintű módosítók típusait is a módosító szófaja szerint megkülönbözteti. Ezeken belül pedig opcionálisan további altípusokat lehet megkülönböztetni. Az SZDT-ben mindegyik reláció szolgál, az ATT.

Az alárendelő mellékmondatok csatolása Az SZDT-ben minden alárendelő mellékmondat az azt tartalmazó tagmondat fejéhez van csatolva (a vonatkozó mellékmondatok is). Ez legalábbis a vonatkozó mellékmondatok tekintetében határozottan nem felel meg az UD-ben megfogalmazott elveknek. Ráadásul semmi nem utal az annotációban arra, hogy a főmondatban sok esetben jelen lévő utalószónak és a hozzá tartozó alárendelő mellékmondatnak bármi köze lenne egymáshoz. Bár az UD elvek szerint az utalószók esetében ennek így is kellene lennie, és valószínűleg inkább expletív névmásokként kellene őket annotálni, végső soron nem lehet megúsni, hogy ha bármire is akarjuk használni az annotációt, összekapcsoljuk megfelelő elemeket egymással. Megoldásunkat és az előbbi *egy-több megfeleltetés*-problémát az 1. ábra illusztrálja. Az ábrán látható, hogy az SZDT ‘all-in-one’ ATT relációja a belőle automatikusan generált UD annotációban számtalan különböző reláció alakját ölti. A ábrán látható rövid mondatban az eredeti mondat ATT típusú dependensei 6 merőben különböző szerepet kapnak: van itt melléknévi jelző (**amod:att**), NP módosító (**nmod:att**), vonatkozó mellékmondat (**acl:relcl**: ezt a konverzió során automatikusan helyesen átkötjük az általa ténylegesen módosított NP fejére), számnévi módosító (**nummod:att**), tárgyi alárendelő mellékmondat (**ccomp:obj**), oblikvuszi alárendelő mellékmondat (**acl:ccomp:obl**: ezt is automatikusan átkötjük a vonzat esetét hordozó megfelelő utalószóra) és birtokos (**nmod:poss**).



1. ábra. Az SZDT ‘all-in-one’ ATT relációja 6 különböző relációvá lényegül át egy röpké mondat UD-beli reprezentációjában.

A relációk iránya és topológiája A relációk feje, illetve a komplexebb szerkezetekhez (pl. a többszörös mellérendeléshez) rendelt annotáció topológiája szempontjából is sok alapvető eltérés van az UD és az SZDT között. Az UD-ben *a tartalmas szó a fej*-elv érvényesül, így pl. a kopulás és a névutós szerkezetek feje is a névszó. A kopulás, illetve a névutós szerkezetek átalakítása általában viszonylag problémamentesen megoldható, a mellérendeléssel kapcsolatos problémákra alább részletesebben kitérünk (2.5).

Üres elemek Az SZDT beszűrt üres elemekkel (**VAN**, **ELL**) operál a zérókopula és az ellipsis tekintetében is. Ez mindenképpen problémát jelent a parserek működése szempontjából, hiszen a nyers szövegből nyilvánvalóan hiányoznak ezek

a testetlen elemek. Az elemzés folyamán valamikor valahogyan a szövegbe kellene kerülniük ahhoz, hogy SZDT-kompatibilis elemzés jöhesse létre. Bár azzal kapcsolatban a magyar beszélőknek világos intuíciója van, hogy hol lenne a zérókopula helye egy konkrét magyar nyelvű tagmondatban, vagy hogy honnan hiányzik egy elliptált elem, a treebankbe ezek az elemek teljesen véletlenszerű helyekre lettek beszúrva, néha nem is abba a tagmondatba, ahova tartoznak. Sajnos az ellipsis jelölése az SZDT-ben nemcsak a törölt elem helyének jelölése tekintetben következtelen: sok ténylegesen elliptikus szerkezetben egyáltalán szerepel az ELL elem.

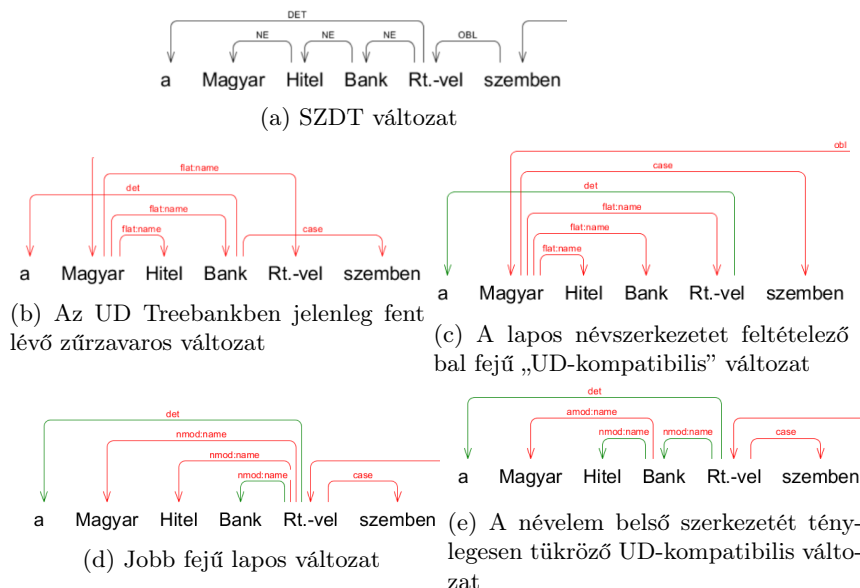
A zérókopula kezelését az UD megnyugtatóan megoldja. Ugyanakkor az elliptikus szerkezetek annotációjára az UD 2-es verziójában elfogadott megoldás (a törölt elem legprominensebb dependensének fejjé való előléptetése, és a többi elem hozzacsatolása az *orphan* reláció használatával) önmagában nem teszi lehetővé a tényleges viszonyok visszafejtését az annotációból. Ráadásul ha csak egy konstituens marad az elliptikus tagmondatban, egyáltalán nem derül ki az annotációból, hogy ellipsisról van szó.

Megoldás A többértelmű relációk konverziójánál az adott token és konstrukciót alkotó egyéb elemek morfológiai/lexikai tulajdonságaira támaszkodva egyértelműsítjük a szerkezetet, és ezután alakítjuk át a megfelelő topológiájúvá az annotációt (fej-dependens viszony megfordítása, az eredeti fej nem-lokális dependenseinek átkötése az új fejre (vagy az összesé a konstrukciótól függően), adott esetben láncolt szerkezetek átalakítása lapos szerkezetté). Az átalakítás az egyértelmű (pl. zérókopulás/névutós/mutató determinánsos) szerkezetek esetében hasonlóan megy. Az alárendelő mellékmondatok átcsatolására készített megoldásunk eredményét az 1. ábrán szemléltettük. A mellékmondatok típusának azonosításához a mellékmondatot tartalmazó mátrixtagmondatban keresünk utalószót (itt számos esetben a mellékmondat kötőszavára támaszkodhatunk, pl. *akkor-amikor*, *annál-minél*, *ott-ahol*, stb.), illetve az utalószóként szóba jöhető deiktikus névmási elem és a tagmondat távolságát és sorrendjét is figyelembe vesszük az alkalmazott heurisztikákban. Vonatkozó mellékmondatok esetében a legközelebbi névszói fejet célozzuk meg, ha utalószóhoz való csatolás nem lehetséges. Tárgyi alárendelő mellékmondatot feltételezünk a *hogy* kötőszós mellékmondatoknál, ha a mátrix ige definit ragozása, és nincs explicit tárgya. Egyébként a megtalált utalószó esete, illetve a kötőszó alapján döntünk a mellékmondat típusáról.

2.4. Névelemek

Az SZDT-ben minden névszerű kifejezés (beleértve a művek stb. címeit is) minden eleme tulajdonnév szófajcímkével van címkézve (a funkciószavak is), és a kifejezés elemei koordináció kezeléséhez hasonlóan láncba vannak fűzve, csak ebben az esetben a lánc jobb fejű (2a ábra). Az UD 2 specifikáció szerint a belső szerkezet és fej nélküli névkifejezéseket a koordinációhoz hasonlóan fixen bal fejű kvázilapos szerkezetként kell ábrázolni (más persze a függőségi viszony, mint a

koordinációnál: (2c ábra)). Mivel a magyarban az NP-k minden esetben a leghatározottabban jobb fejűek, ezért ez a javasolt annotáció ilyen formában biztosan nem használható (pl. a név esetét adó névutó, amely maga instrumentalist vonz, az alanyesetben álló első szóhoz kapcsolódik, ami nonszensz), helyette például valamilyen jobb fejű lapos annotáció használata lenne indokolt (2d ábra)). Jelenleg az UD Treebankben egyébként a 2b ábrán látható teljesen zűrzavaros szerkezet szerepel. Ezt feltehetőleg az UD 1.0-s változatban szereplő eredetileg helyesen jobb fejű szerkezeteket tartalmazó annotáció programozott „megrongálásával” hozta létre az UD treebankek valamelyik magyarul nem tudó adminisztrátora. Ugyanakkor az UD 2 specifikáció szerint a világos belső szintaktikai szerkezettel és függőségi viszonyokkal rendelkező névelemek annotációjának ezeket a viszonyokat kellene tükröznie (2e ábra)).



2. ábra. Egy névelem lehetséges (d,e) (és nemkívánatos (b,c)) ábrázolásmódjai

A világos belső szintaktikai szerkezettel rendelkező nevek, pl. a művek címeinek egy része a magyar esetében egyébként problémát jelent abból a szempontból is, hogy míg maga a cím jobb fejűnek tekinthető (lévén kívülről nézve egy NP), a belső szerkezet feje egész más lehet (pl. egy ige), és így a jobb szélén álló elem akár egyszerre két esetben állhat (6a,6c). Tulajdonképpen ezekben az esetekben egyfajta lexikai ellipszisről van szó (6b,6d).

- (6) a. Stohl András párja lesz Éder Enikő a *Hegedűs a háztetőn*-ben.
 b. Stohl András párja lesz Éder Enikő a *Hegedűs a háztetőn* című musicalben.
 c. A HBO Max levette az *Elfújta a szél/szelet*.
 d. A HBO Max levette az *Elfújta a szél* című filmet.

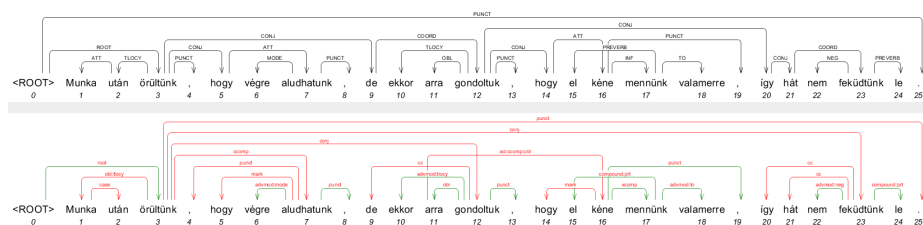
Megoldás Az névelemek szétválogatása és a fenti szempontok szerinti újraelemzése mindenképp jelentős erőfeszítést igényel. Az ehhez szükséges munkába egyelőre nem vágunk bele: a jelenleg változatban a jobb fejű lapos annotációnál maradtunk. A típusok szintjén a morfológiai elemző elemzéseitől való eltérések fele a névelemek annotációjából adódik.

2.5. Mellérendelés

A mellérendelésről azt tanultuk, hogy exocentrikus szerkezet, és mint ilyen alapvetően problematikus a kizárólag endocentrikus szerkezetekben gondolkodó függőségi grammatika számára. Ezért bármit is teszünk, az annotáció mindenképp önkényes lesz, különös tekintettel arra, hogy a mellérendelésnek kénytelen lesz feje és iránya lenni. UD specifikációjában szereplő tartalmas fejeket összekötő kvázilapos szerkezet (minden mellérendelt elem közvetlenül az önkényesen kijelölt fejhez van kötve) elvileg kevesebb problémát okoz, mint az SZDT-ben alkalmazott vegyes láncba fűzött és a mellérendelt elemet a kötőszón keresztül csatoló megoldás. Az előbbiben ugyanis minden mellérendelt konstituens feje maximum egy lépés távolságra van a mellérendelő szerkezet önkényesen kijelölt fejétől, és így determinisztikus módon elérhetők mindegyik elemből a szükséges információk, míg a második megoldásban nem korlátos az adott koordinált elem és a valódi grammatikai szerepére vonatkozó információk helye közötti távolság.

Egyik módszer sem teljesen alkalmas a különböző egymásba ágyazott szerkezetek megkülönböztetésére. Az UD-ben alkalmazott reprezentáció esetében az A,B,C és az (A,B),C szerkezethez tartozik azonos fa, az A,(B,C) szerkezeté különbözik tőlük. Az SZDT-ben alkalmazott láncolt megoldásnál az első és a harmadik ad azonos szerkezetet, a második különbözött. Ezért nem lehet az SZDT-ben alkalmazott láncolt szerkezeteket ‘ész nélkül’ az UD specifikációban ajánlott szerkezetekké alakítani. A 3. ábrán látható, hogy a valójában A,(B,C) típusú szerkezetet ábrázoló SZDT annotációt A,B,C szerkezetűnek feltételezve és azt az UD specifikáció által javasolt formára alakítva az ábrán alul szereplő annotációt kapjuk, amiről azt a badarságot olvashatjuk le, hogy a mondat szerint *Munka után örültünk, hogy végre aludhatunk, így hát nem feküdtünk le*. A korpuszban a kettőnél több tagmondatot tartalmazó konkrét tagmondatmellérendelés-példákat nézegetve nagyon hamar arra a következtetésre juthatunk, hogy az esetek nagyobb részében teljes zagyvaság jön ki, ha hozzányúlunk a tagmondatok közötti viszonyokhoz és A,B,C mellérendelést feltételezve átkötögetjük a harmadik tagmondatot az elsőre. Az SZDT-ben szereplő mellérendelés-annotáció

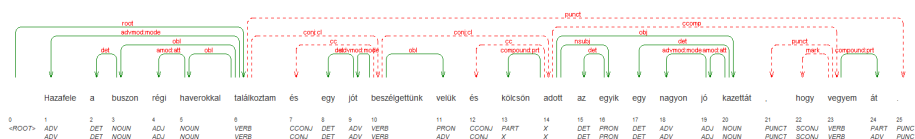
általában jól tükrözi a diskurzusszerkezetet, a tagmondatok közötti időbeli, illetve a logikai viszonyokat. Általában ténylegesen az előző tagmondathoz csatlakozik a következő mondanivalója, ha meg nem, azt ez a típusú annotáció is jól ki tudja jelezni az (A,B),C szerkezet megadásával. Tagmondatok esetében, úgy tűnik, nem sok értelme van többszörös mellérendelésről beszélni, mert még explicit kötőszó hiányában is gyakran legalábbis az események sorrendjét tükrözi a tagmondatok sorrendje. Ezért nincs értelme a tagmondat-mellérendelések annotációjába belepiszkálni.



3. ábra. Mi lesz, ha az SZDT mellérendelt tagmondatait ‘ész nélkül’ az UD specifikáció szerinti formára hozzuk: *Munka után örültünk, hogy végre aludhatunk, így hát nem feküdtünk le.*

A névszói kifejezések koordinációjánál ugyanakkor megfontolandó lehet az UD specifikációban javasolt mellérendelés-annotációra való áttérés. Az UD és az SZDT annotáció különös közös fogvatékossága, hogy a frázis- és tagmondatkoordináció nincs megkülönböztetve: ugyanazt a relációt használják a két viszony megjelölésére. Ez például annak az egy szintaktikai annotációval szemben támasztható viszonylag alapvetőnek tűnő elvárásnak a teljesíthetőségét is megkérdőjelezi, hogy a tagmondathatárokat az annotáció alapján meg lehessen állapítani. Ugyanakkor az UD specifikáció megengedi az alaprelációkon belül altípusok létrehozását, tehát megkülönböztethetünk egy *conj:ph* frázis-mellérendelés és egy *conj:c1* tagmondat-mellérendelés relációt. Elvileg az összekapcsolt fejek kategóriája alapján meg lehet különböztetni a kettőt: a tagmondatok feje általában ige. Az UD annotációban azonban a névszói állítmányok és az ellipsis esetében ez nincs így, bár az utóbbiaknál szerencsés esetben egy *orphan* reláció jelenlétéből meg lehet állapítani, hogy ellipszist tartalmazó tagmondatról van szó (ha egynél több konstituens maradt ‘árván’ az elliptikus tagmondatban).

Az SZDT annotáció esetében az okoz problémát, hogy az ellipsis időnként nincs jelölve, illetve hogy a ‘hibás szó’ szófajjelölést kapott szavak szófaját nem lehet az annotációból kiolvasni. Szerencsés esetben azonban legalább a koordinációt kifejező *conj* reláció fejének annotációja használható, ami alapján mégis meg tudjuk különböztetni a frázis- és a tagmondatkoordinációt (így azonosíthatjuk, hogy a 4 ábrán látható mondatban a *beszélgettünk* és a hibásnak megjelölt (*kölcsön*) *adott* közötti viszony *conj:c1*), azaz tagmondat-mellérendelés.



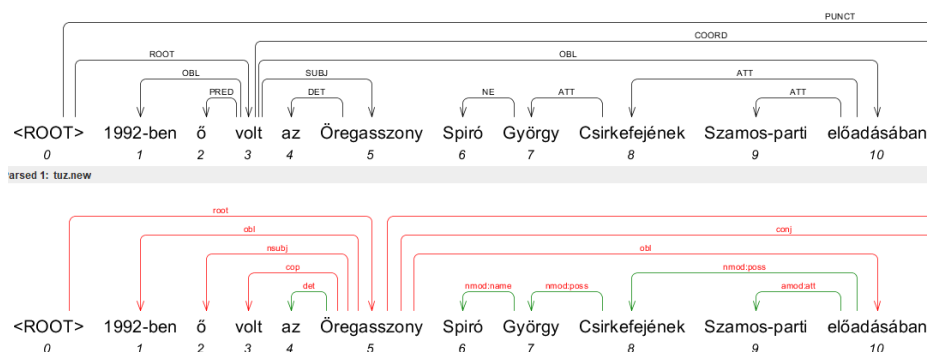
4. ábra. Tagmondat-mellérendelés azonosítása hibásnak annotált tagmondatfej esetén (*kölcsön adott*).

2.6. Egyéb annotációs hibák és inkonzisztencia

Alany-állítmány tévesztés Az SZDT-ben a harmadik személyű **névszói állítmányt** tartalmazó tagmondatok annotációjában az alany és az állítmány szinte minden esetben fel van cserélve olyankor, ha az alany fókuszos, és az állítmány definit. Az elv láthatóan az volt, hogy ha két definit harmadik személyű NP verseng az alany és az állítmány szerepéért, akkor a fókuszos elemet jelölték állítmánynak (5. ábra). A fókusz és az állítmány azonban különböző dolog. A zavart az okozza, hogy fókuszos mondatok használatának szerencsefeltétele az, hogy tudjuk, hogy az adott állítás (az állítmány) valakire/valamire igaz, a mondatban az új információ az, hogy kire/mire (ez tűnik tehát állításnak). Az alany legnyilvánvalóbb ismérve, hogy mondatban a finit ige mindig az alannal van egyeztetve (nyilván akkor is, ha az fókuszban áll). A kopulától különböző igét tartalmazó mondatokban fel sem merül, hogy a fókuszban álló elemet állítmánynak tekintjük. Minden esetben gyanús, ha névmás van állítmánynak jelölve. Ha az állítmánynak jelölt névmást (vagy egyéb gyanús névszói elemet) első/második személyű névmásra tudjuk cserélni, és a mondat értelmes és lényegében hasonló jelentésű marad (és persze a finit ige ilyenkor első/második személyű egyeztetésre vált), akkor egyértelműen hibás volt az eredeti annotáció.

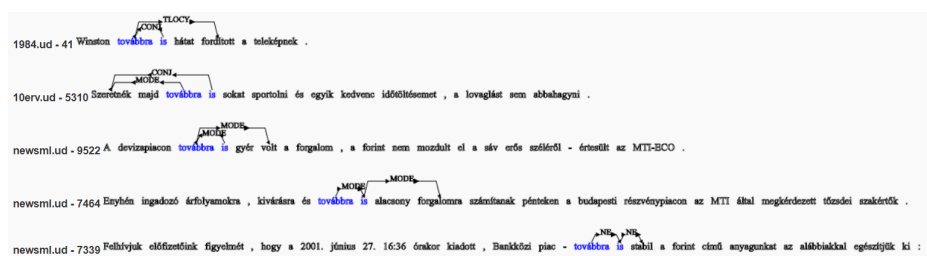
Megoldás A gyanús névszói állítmányokra (amelyeknél sokkal valószínűbb, hogy referáló kifejezés fejeként szerepel) rákeresve találtunk rá az ilyen típusú hibásan annotált mondatokra. Kézzel jelöltük meg a valóban tévesen annotált esetekben a kopulát, és ezekben a mondatokban automatikusan megcseréltük az alanyt és az állítmányt.

Inkonzisztens annotációk – ismétlődő mondatok Az SZDT-ben 1801 mondat(token) egybeesik valamelyik másik korpuszbeli mondattal (ismétlődik). Néhány esetben egy szövegrész véletlen ismétlődéséről van szó (itt az ismétlődő részt törölni kellene), de az esetek nagyobb részében nem ez a helyzet. Az 1801 ismételt mondat(token) 578 különböző mondattípust testesít meg, ezek közül 440 mondat kétszer fordul elő, és van olyan mondat is, ami ötvenkétszer (*Részletek a hétfői Befektetőben.*). 118 ismétlődő mondattípusnak két különböző elemzése van, 8-nak három.



5. ábra. Az SZDT-ben hibásan felcserélt alany-állítmány viszony javítása a mondat UD-beli reprezentációjában.

Inkonzisztens annotációk – ismétlődő szósorok Emellett a korpusz 27245 olyan legalább kételemű szótokensorozatot tartalmaz, amelyen belül a függőségi viszonyok legalább két különböző módon vannak annotálva, illetve 13815 olyan sorozatot, ahol a fő szófajcímek sorozata nem azonos. Természetesen az esetek nagy részében valódi többértelműségeknek felelnek meg az annotációs különbségek, illetve az inkonzisztenciák nagy része a zérókopula és az ellipszis annotációjára használt üres tokenek teljesen következtelen helyekre történő beszúrásából ered. Ezeknek az inkonzisztenciáknak a feltárásához az *Errator* eszköz egy módosított verzióját használtuk (Wisniewski, 2018). A *továbbra is* szókapcsolat annotációjával kapcsolatos valódi inkonzisztenciákra mutatunk példákat a 6. ábrán, amely jól érzékelteti egyrészt az *is* klitikum és az egyéb funkciószavak szófaji besorolásával és szintaktikai kapcsolásával kapcsolatos bizonytalanságokat.



6. ábra. Példák a *továbbra is* szókapcsolat annotációjával kapcsolatos inkonzisztenciákra az SZDT-ben

Megoldás Az ismétlődő mondatok inkonzisztens annotációja kézzel átnézhető. A rövidebb ismétlődéseknél feltárt problémák kezelésénél a gyakori típusokat

(pl. az *is* eseteit) programozott megoldással javíthatjuk. Az egyedibb esetekben valószínűleg szintén csak a kézi javítás segít.

3. Összefoglalás

Cikkünkben a Szeged Dependencia Treebanket UD-kompatibilis verziójának előállítására tett erőfeszítéseink során felmerülő problémákat, illetve az ezek megoldására tett kísérleteinket mutattuk be. Az automatikusan megoldható átalakítások nagy részét implementáltuk, és sok manuális javítást, illetve ellenőrzést is elvégeztünk. Azonban a feladat elég sok élő munkát igényel, így egyelőre folyamatban lévő munkálatról beszélhetünk. Az egyelőre félkész terméket (illetve majd remélhetőleg a „kész” változatot) az `nlpg.itk.ppke.hu/resources` címen tesszük elérhetővé.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással az FK 125217 és a PD 125216 számú projekt keretében az FK 17 és a PD 17 pályázati program, valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatásával a 2018-1.2.1NKP-2018-00008 azonosítójú projekt keretében valósult meg.

Hivatkozások

- Dozat, T., Qi, P., Manning, Ch.D.: Stanford’s graph-based neural dependency parser at the CoNLL 2017 shared task. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 3-4, 2017. pp. 20–30 (2017), <https://doi.org/10.18653/v1/K17-3002>
- Nivre, J., de Marneffe, M.C., Ginter, F., Hajič, J., Manning, Ch.D., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D.: Universal Dependencies v2: An evergrowing multilingual treebank collection. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4034–4043. European Language Resources Association, Marseille, France (May 2020), <https://www.aclweb.org/anthology/2020.lrec-1.497>
- Novák, A., Laki, L.J., Novák, B., Dömötör, A., Ligeti-Nagy, N., Kalivoda, Á.: Egy magyar nyelvű kérdezőrendszer. In: XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 83–95 (2019)
- Novák, A., Rebrus, P., Ludányi, Zs.: Az emMorph morfológiai elemző annotációs formalizmusa. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). pp. 70–78 (2017)

- Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria (2013)
- Vincze, V., Simkó, K., Szántó, Zs., Farkas, R.: Universal Dependencies and morphology for Hungarian - and on the price of universality. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 356–365. Association for Computational Linguistics, Valencia, Spain (Apr 2017), <https://www.aclweb.org/anthology/E17-1034>
- Vincze, V., Szauder, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian dependency treebank. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (szerk.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
- Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholcz, I., Novák, A., Indig, B., Prószéky, G., Farkas, R., Vincze, V.: Az e-magyar digitális nyelvfeldolgozó rendszer. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). pp. 49–60 (2017)
- Wisniewski, G.: Errator: a Tool to Help Detect Annotation Errors in the Universal Dependencies Project. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (szerk.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)
- Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., Petrov, S.: CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. pp. 1–21. Association for Computational Linguistics, Brussels, Belgium (October 2018), <http://www.aclweb.org/anthology/K18-2001>

Értsük meg a magyar entitásfelismerő rendszerek viselkedését!

Farkas Richárd¹, Nemeskey Dávid Márk², Zahorszki Róbert¹, Vincze Veronika³

¹Szegedi Tudományegyetem, Informatika Intézet

²Eötvös Lóránd Tudományegyetem, Digitális Bölcsészet Központ

³MTA-SzTE Mesterséges Intelligencia kutatócsoport

rfarkas@inf.u-szeged.hu

Kivonat: A nyelvtechnológiai megoldásokat hagyományosan egy valós életből származó szöveghalmaz tanító és tesztadatbázisra bontott verzióján szokás kiértékelni, e módszer azonban több buktatóval is rendelkezik. A CheckList egy új-fajta kiértékelési módszertan, mely különböző nyelvi jelenségeket definiál, továbbá az egyes jelenségekre külön tesztkörnyezeteket állít fel, melyek az adott alkalmazás viselkedését hivatottak tesztelni. Ebben a tanulmányban a magyar nyelvű névelem-felismerési (NER) feladatra alkalmazzuk a CheckList módszertanát. Ehhez 9 nyelvi jelenséget¹ definiálunk, mondatsablonokon keresztül 27 tesztkörnyezetet állítunk fel és három magyar névelem-felismerő rendszert értékelünk ki a CheckList módszertanában. Elemzésünk megmutatja, hogy ez a módszertan közelebb visz minket ahhoz, hogy megértsük a magyar entitásfelismerők viselkedésének megértését.

1 Bevezetés

Az elmúlt évtizedekben a nyelvtechnológiai megoldásokat szinte minden esetben egy valós életből származó szöveghalmaz, kézzel jelölt, tanító és kiértékelő adatbázisra válogott verzióján értékelték ki. A tanító adatbázison gépi tanult rendszerek (vagy az alapján kézzel épített szabályrendszereket) pontosságát a kiértékelő adatbázison mérjük meg és ezt egyetlen számmal (pl. accuracy, F1-érték vagy BLEU score) írjuk le. Az elérhető adatbázisokon mindig verseny indul, és aki a kiértékelési metrikában akár csak fél százalékponttal jobb eredményt ér el, mint a korábban publikált legjobb eredmény, az már publikálható eredménynek számít. Ezt a tudománytörténeti jelenséget *leaderboard paradigmának* nevezi Ethayarajh és Jurafsky (2020). A leaderboard paradigma számos problémát vet fel, amelyek orvoslására az elmúlt két-három évben több javaslat is megjelent a *ACL és EMNLP konferenciákon.

¹ Az eredeti CheckList módszertanban használt “capability” fogalmát ‘nyelvi jelenségeként’ fordítjuk, elkerülve a ‘nyelvi képesség’ fogalmának túlterhelését, mivel utóbbit a magyar nyelvészeti szakirodalom főként a nyelvsajátítás és idegennyelv-tanulás területein alkalmazza (a ‘linguistic ability’, illetve a ‘language skill’ megfelelőjeként).

A kiértékelő adatbázisokon való kiértékelés természetesen hasznos, de mivel az adatbázis eloszlását követi, ezért számos torzítást tartalmazhat - például mert egy szűk téma, zsáner vagy stílus dominálja -, sokszor az adott modell túláltalánosít a tanító adatbázis alapján és egy másik kiértékelő adatbázison már kevésbé jó eredményt ad. Továbbá, ha egyetlen számmal írjuk le a rendszer teljesítményét, akkor abból nem tudjuk meg, hogy hol és miért hibázik a rendszerünk, azaz nem értjük meg, hogy hogyan viselkedik a vizsgált rendszer. Ez pedig elengedhetetlen ahhoz, hogy egy rendszer, egy konkrét valós életbeli feladatra való alkalmazhatóságáról dönteni tudjon az alkalmazás-fejlesztő.

Az egyes feladatok megoldásához számos nyelvi jelenség kezelésére szükség lehet, és ezek gyakorisága és fontossága igencsak eltérő lehet egymástól. Megeshet, hogy az adott módszer a legfontosabb, legalapvetőbb példákon jól teljesít, azonban a nehezebb, bonyolultabb példákon elbukik, vagy esetleg ennek a fordítottja: néhány alapvető példát elront, de a nehezebbeken jól teljesít (például mert ezek túl vannak reprezentálva a tanító adatbázisban), a számszerű eredményekben azonban e különbségek nem mutatkoznak meg.

A fenti problémák kiküszöbölésére Ribeiro és mtsai (2020)² bevezették a „CheckList” tesztelés fogalmát, melyet részben a szoftverfejlesztésben használatos tesztelési módszertan inspirált. A CheckList egy újfajta kiértékelési módszertan, mely különböző nyelvi jelenségeket definiál, amelyeket a rendszernek az adott feladat (és nem adatbázis!) megoldásához bizonyítani kell. Az egyes jelenségekre külön tesztkörnyezeteket állít fel, melyek az adott alkalmazás viselkedését hivatottak tesztelni. Ez a fajta diagnosztikus tesztelés jól kiegészíti a kiértékelő adatbázison számolt metrikákkal kapott minőségellenőrzést.

Ribeiro és mtsai (2020) az angol nyelv vonatkozásában mutatják be módszerüket a szentimentelemzés, duplikált kérdések azonosítása és a gépi szövegértés területére alkalmazva. Ebben a tanulmányban a magyar nyelvű névellem-felismerési (NER) feladatra alkalmazzuk a CheckList módszertanát. Ehhez 9 nyelvi jelenséget³ definiálunk, mondatsablonokon keresztül 27 tesztkörnyezetet állítunk fel és három magyar névellem-felismerő rendszert értékelünk ki a CheckList módszertanában. Elemzésünk megmutatja, hogy ez a módszertan közelebb visz minket ahhoz, hogy megértsük a magyar entitásfelismerők viselkedésének megértését.

2 Kapcsolódó munkák

Az elmúlt néhány évben számos munka kérdőjelezi meg a nyelvtechnológiai kutatások leaderboard paradigmáját (Ethayarajh és Jurafsky, 2020). Ethayarajh és Jurafsky (2020) a végfelhasználói alkalmazások fejlesztőinek (NLP practitioners) szempontjából tárgyalja, hogy a pontosság metrikák mellett milyen szempontok fontosak egy feladatra adott megoldás szempontjából. Például javasolja a futásidők és energiafelhasználás (Green AI) feltüntetését minden publikációban, hiszen a valós életben, ha két modell közül az egyik néhány százalékponttal pontosabb, de nagyságrendekkel erőforrásigényesebb,

² ACL 2020 best paper

³ Az eredeti CheckList módszertanban használt “capability” fogalmát fordítjuk ‘nyelvi jelenségként’.

mint egy másik modell, akkor az alkalmazók a valamivel pontatlanabbat fogják preferálni. Egy másik fő kritika a tanító- és kiértékelő adatbázisokon való mérésekkel szemben a *robustusság* megismerésének hiánya, ugyanis egyetlen adatbázison kiértékelve, nem tudjuk, hogy a rendszerek mennyire jól viselkednek a tanító adatbázis eloszlásától eltérő példákon, mennyire tűrik a bemenet változásait, illetve mennyire torzítanak egyes demográfiai tulajdonságok irányába (ML fairness).

Ebben a munkában, a robustusság témakörébe tartozó CheckList (Ribeiro és mtsai, 2020) kiértékelési módszertant használjuk. A CheckList az úgynevezett *black box diagnosztikus tesztek* közé sorolható, hiszen célja annak felmérése, hogy hol és miért hibázik a tesztelt rendszer, valamint feltesszük, hogy a rendszer belső működéséhez nem férünk hozzá, az fekete dobozként - egy bemenetre visszaad egy eredményt - áll rendelkezésre (Paroubek és mtsai, 2007).

A CheckList egy általánosított keretrendszert ad nyelvtechnológiai alkalmazások különböző viselkedési tesztjének definiálására. Például invariancia típusú tesztekkel tudjuk a zajjal - például elírásokkal - szembeni robustussági teszteket definiálni, vagy más típusú tesztekkel tudjuk a rendszer logikai konzisztenciáját tesztelni. A CheckList kifejezetten végfelhasználói nyelvtechnológiai alkalmazások kiértékelést célozza meg, és olyan eszközt ad, amit a nyelvtechnológiában járatlan, de az adott célalkalmazás szakértő felhasználói is tudnak használni. Ez fontos különbség egyéb javaslatokkal szemben. Például a köztes modulok robustusságának kiértékelésére használt extrinziкус tesztelés - amikor a modult különböző ráépülő alkalmazásoknak nyújtott hasznosság szerint értékeljük - nem alkalmas célalkalmazások tesztelésére.

A CheckList célkitűzése, hogy megértsük a black box rendszer viselkedését. Ebben az aspektusban a megmagyarázható MI (eXplainable AI) tárgykörébe is besorolható. Ezen algoritmusok közül is kiemelkedik azonban egyszerűségének és univerzalitásának köszönhetően. Például minden feladathoz más és más interpretációs algoritmusokra van szükség (Arrieta és mtsai, 2020), míg a CheckList keretrendszerben bármilyen feladatot kiértékelhetünk. Hasonlóan a neurális modellek megértését célzó ún. próbák módszere (probes) is minden nyelvi jelenség tesztelésére külön algoritmust követel meg (Hewitt és Manning, 2019), míg a CheckListtel bárki tesztelhet bármilyen nyelvi jelenséget.

3 Magyar NER checklist

Figyelembe véve a magyar nyelv tulajdonságait és a névelem-felismerésben fontos nyelvi jegyeket, összeállítottunk egy olyan nyelvi tesztsorozatot, mely segítségével célirányosan tudjuk tesztelni a NER-rendszerek teljesítményét, továbbá meg tudjuk állapítani, mik az egyes rendszerek erősségei és gyenge pontjai. Alább bemutatjuk e jellemzőket, valamint az egyes tesztípusokat.

3.1 Tesztípusok

Minimális működés tesztje (Minimum Functional Test, MFT): Azon alapvető példák tartoznak ide, melyeknek helyes kezelését elvárjuk egy tulajdonnév-felismerő

rendszerrel. Például az *X. Y., Magyarország köztársasági elnöke* példában bármi/bárki is álljon *X.Y.* helyén, az személynév (PERSON) címkét kell hogy kapjon.

Invariancia (INV): Ha megváltoztatjuk bizonyos módon a bemeneti mondatot, az nem okozhat változást a rendszer predikciójában. Például egy szórendi csere általában nem befolyásolja a címkézést (*London_{LOC} mellett ülésezett a NOB_{ORG}* vs. *A NOB_{ORG} London_{LOC} mellett ülésezett*).

Elvárt változás (DIR): Ribeiro és mtsai (2020) definálnak egy harmadik típusú tesztet is, ahol a bemenet változtatásával a predikció irányának megváltozását tesztelik. Az eredeti definíció alapján, az INV és DIR tesztet el lehet végezni jelöletlen szövegeken is - míg az MFT-hez annotált példák szükségesek -, hiszen ezeknél csak azt vizsgáljuk, hogy megváltozik a predikció, és azt nem teszteljük, hogy az eredeti szövegen helyes volt-e ez a predikció. Ribeiro és mtsai (2020) erre egy szentiment elemzési példát hoznak, ahol egy negatív töltetű mondatral kiegészítve a szöveget, elvárjuk, hogy a pozitív osztály valószínűsége ne növekedjen. Mivel a tesztelt magyar NER rendszereink alapesetben nem adják meg az egyes címkék valószínűségét, ezért igazi DIR típusú tesztet nem használunk jelen munkában. Megjegyezzük azonban, hogy magyar névelem-felismeréshez is lehet olyan változásokat eszközölni kontrollált - azaz kézzel címkézett esetekben - ahol a bemenet változtatásától egy entitás osztályának megváltozását várjuk el. Például ha egy helynév névelőt kap, akkor bizonyos kontextusokban szervezetrévé lesz belőle:

Manchesterben_{LOC} játszott Ronaldo_{PER} vs. *A Manchesterben_{ORG} játszott Ronaldo_{PER}*

Mivel ez az elvárt változás nem felel meg pontosan az eredeti DIR definíciónak, ezért az ilyen jellegű tesztet MFT-ként fogalmazzuk meg, a fenti példából két darab MFT teszt lesz:

Egy helynév: *Manchesterben_{LOC} játszott Ronaldo_{PER}*

Névelős helynév: *A Manchesterben_{ORG} játszott Ronaldo_{PER}*

3.2 Nyelvi jelenségek

A magyar nyelv morfológiailag gazdag volta miatt több morfológiai, illetve szintaxis alapú nyelvi jelenségre is építhetünk a névelem-felismerés hatékonyságának tesztelése terén. Ezek mellett néhány szemantikai jellegű változásra épülő jelenséget is bemutattunk.

Szókins: Az adott tulajdonnévre jellemző legtipikusabb szókészletet reprezentáló példamondatok tartoznak ide, például: *A szomszédomat Fekete_{B-PER} Péternek_{L-PER} hívják.*

Névelő: Ha névelőt kap az adott tulajdonnév, akkor adott irányú változást mutat (vagy nem mutat változást) a címkézésben, például: *Fordnál_{PER} járt a szépségkirálynő* vs. *A Fordnál_{ORG} járt a szépségkirálynő.*

Toldalékolás: Eltérő toldalékolás (pl. esetrag) esetén adott irányú változást mutat (vagy nem mutat változást) az adott tulajdonnév, például: *A Gyulában_{ORG} focizott Feri* vs. *A Gyulával_{PER} focizott Feri.*

Névutó: A névutó cseréje esetén adott irányú változást mutat (vagy nem mutat változást) az adott tulajdonnév, például: *London_{LOC} mellett ülésezett a NOB_{ORG}* vs. *London_{MISC} után ülésezett a NOB_{ORG}*.

Többes szám: Ha az adott tulajdonnevet többes számba tesszük, adott irányú változást mutat (vagy nem mutat változást) a címkéje, például: *Az autóversenyt Ford_{PER} nyerte* vs. *Az autóversenyt Fordok_{MISC} nyerték*.

Predikátum cseréje/szemantikai szerepek változása: Más predikátum esetén adott irányba változik (vagy változatlan marad) az adott tulajdonnév címkéje, például: *A cég felvásárolt még egy gyárat a Mercedes_{ORG} mellett* vs. *A cég megvásárolt még egy telket a Mercedes_{LOC} mellett*.

Taxonómia: Szinonimák, antonimák, hipernimák stb. cseréje esetén adott irányú változást mutat (vagy nem mutat változást) az adott tulajdonnév, például: *A Manchesterben_{ORG} futballozott Ronaldo_{PER}* vs. *A Manchesterben_{ORG} játszott Ronaldo_{PER}*.

A fenti nyelvi jelenségek mellett külön megvizsgáltuk azokat az eseteket is, amikor többtagú tulajdonneveket kell azonosítani, valamint a magyar nyelv szórendi jellemzői miatt külön figyelmet fordítottunk azokra az esetekre is, amikor szórendi okok miatt két azonos típusú, ámde különálló névelem került egymás mellé. Lásd az alábbi példákat:

Többtagú tulajdonnevek: *Megalakult a Magyar_{B-ORG} Nemzeti_{I-ORG} Bank_{I-ORG}*.

Egymást követő azonos típusú tulajdonnevek: *Évi_{B-PER} Pétertől_{B-PER} egy könyvet kapott*.

3.3 Nyelvi variációk

A fenti nyelvi jelenségeken túl a CheckList módszertana lehetőséget ad arra is, hogy további variációs lehetőségeknek vessük alá a tesztmondatainkat. Míg Ribeiro és mtsai (2020) a tagadást és szórendi variációkat szintén nyelvi jelenségeként kezelik, mi a magyar nyelv sajátosságai miatt indokoltabbnak látjuk e két variációs lehetőséget külön-külön alkalmazni a nyelvi jelenségekre. Így tehát a fenti jelenségeket megháromszorozhatjuk, a fent felsorolt alapesetek mellett beszélhetünk tagadott változatokról és szórendi variánsokról is, amelyek szintén elvárt viselkedést támasztanak a tulajdonnévi címkék esetében. Példaként véve az egyik névutós teszt esetet, itt minden címke változatlan marad az alapesethez képest:

Alapeset:

London_{MISC} után ülésezett a NOB_{ORG}.

Tagadott variációk:

London_{MISC} után nem ülésezett a NOB_{ORG}.

Nem London_{MISC} után ülésezett a NOB_{ORG}.

London_{MISC} után nem a NOB_{ORG} ülésezett.

Szórendi variációk:

A NOB_{ORG} London_{MISC} után ülésezett.

London_{MISC} után a NOB_{ORG} ülésezett.

A NOB_{ORG} ülésezett London_{MISC} után.

4 Magyar NER checklist kiértékelés

4.1 A tesztadatbázis létrehozása

Mindegyik nyelvi jelenségre kézzel állítottunk össze példamondatokat és sablonokat, melyekre aztán kiterjesztettük a nyelvi variancia szintjeit is, így 9×3 mondatcsoportot kaptunk, melyeken tesztelni tudjuk a NER-modellek viselkedését. A kézzel összeállított sablonokból automatikusan generáltuk a tesztmondatokat, összesen 14649 mondatot és 125442 tokent eredményezve.

4.2 Tesztelt névelem-felismerő rendszerek

Három magyar nyelvű névelem-felismerő rendszert választottunk⁴ és értékeltünk ki ebben a munkában:

- A SzegedNER egy klasszikus jellemzőkinyerésen alapuló Conditional Random Field (CRF) névelem-felismerő (Szarvas és mtsai, 2006a)
- Az mBERT (többnyelvű BERT) egy, a 104 legnagyobb Wikipédián tanított, többnyelvű BERT-Base model (Devlin és mtsai, 2018). A BERT egy kétirányú nyelvmoddellen alapuló ún. kontextualizált szóbeágyazás, ami az egyes szavakhoz egy kontextusfüggő jellemzővektort rendel. A modellt az emBERT könyvtárral finomhangolták névelem-felismerésre (Nemeskey, 2020a).
- A huBERT egy magyar BERT modell, amit a Webcorpus 2.0-n és a magyar Wikipédián tanítottak (Nemeskey, 2020b). Mérete (a szótár kivételével) megegyezik az mBERT-ével, viszont kizárólag magyar szövegeken lett előtanítva, ezért kapacitása nem oszlik szét több nyelv között.

Mindhárom névelem-felismerő a Szeged NE (Szarvas és mtsai, 2006b) teljes korpuszán lett betanítva.

4.3 Kiértékelési metrikák

A 1. táblázat tartalmazza az egyes teszteken, az egyes névelem-felismerő rendszerek hibaarányát (százalékban). A hibaarány pontos definíciója:

- MFT típusú tesztek esetén, azt mérjük, hogy kitüntetett frázisokat milyen arányban címkéz helytelenül a névelem-felismerő. Ennek mérésére, a névelem-felismerésben elfogadottan használt, frázisszintű kiértékelő szkriptet használunk és $\text{hiba_arány}_{\text{MFT}} = 1 - \text{micro_fedés}$

⁴ Nem volt célunk az összes magyar tulajdonnév-felismerő rendszer vizsgálata, viszont a szóbeágyazás-alapú és klasszikus jellemzőkinyerés alapú rendszereket össze akartuk hasonlítani.

ahol `micro_fedés` az egyes névelem osztályok fedésének (recall) súlyozott átlaga.

- INV típusú tesztek esetén, azt mérjük, hogy kitüntetett frázisoknál milyen arányban változik meg a predikció, ha az alaphoz képest tagadást vagy szórendi változtatásokat hajtunk végre. Megjegyezzük, hogy itt az a hiba, ha megváltozik a címkézés (sérti az invariancia elvárását), függetlenül attól, hogy egyébként az alap mondatban helyes vagy helytelen volt-e a predikció. Azaz az is INV hibának számít, ha az alap mondatban helytelen címkézés, míg a módosított mondatban helyes a címkézés, hiszen változás történt.

Megjegyezzük, hogy ezeknél a teszteseteknél félrevezető a konkrét értékeket vizsgálni vagy összehasonlítani, hiszen a példamondat-sablonokon nagyon sok múlik. Konkrét értékek helyett csak a nagyságrendeket érdemes nézni, azaz, hogy átment-e vagy elbukott az adott teszten az adott rendszer.

5 Eredmények

Az 1. táblázat tartalmazza a három rendszer kilenc nyelvi jelenségen elért eredményeit. A kísérletek megismételhetősége kedvéért, a tesztmondat-sablonok, a generáló és kiértékelő szkriptek⁵ elérhetőek a [www.github.com/szegedai/hun_ner_checklist](https://github.com/szegedai/hun_ner_checklist) oldalon.

1. táblázat. Hibaarányok százalékban kifejezve (minél kisebb, annál jobb).

Nyelvi jelenség	Nyelvi variációk	SzNER	mBERT	huBERT	példa
Névelő	MFT	58	58	33	A Manchesterben játszott Ronaldo.
	INV tagadás	20	10	7	A Manchesterben nem játszott Ronaldo.
	INV szórend	60	21	4	Ronaldo a Manchesterben játszott.
Toldalékolás	MFT	45	45	20	A Hamburggal játszott Messi.
	INV tagadás	40	14	15	A Hamburggal nem játszott Messi.
	INV szórend	58	32	4	Messi a Hamburggal játszott.
Névutó	MFT	49	38	33	Rio után ülésezett a MOB.
	INV tagadás	16	0	0	Rio után nem ülésezett a MOB.
	INV szórend	51	6	0	A MOB Rio után ülésezett.

⁵ Ribeiro és mtsai (2020) egy tesztelő felhasználói felületet is implementáltak (<https://github.com/marcotcr/checklist>). A munkánk megkezdésekor úgy tűnt, hogy egyszerűbb saját szkripteket implementálnunk, mint integrálni mindent a checklist eszközbe. A munka végére ebben elbizonytalanodtunk, ezért a jövőben tesztelni tervezzük magát a checklist felhasználói felületet is.

Többes szám	MFT	74	75	47	Fordok nyerték az autóversenyt.
	INV tagadás	48	30	8	Nem Fordok nyerték az autóversenyt.
	INV szórend	48	38	8	Az autóversenyt Fordok nyerték.
Predikátum cseréje	MFT	54	54	23	A Madridban énekelt Beckham.
	INV tagadás	33	2	0	Nem a Madridban énekelt Beckham.
	INV szórend	36	20	4	Énekelt Beckham a Madridban.
Taxonómia	MFT	53	53	26	A Madridban focizott Beckham.
	INV tagadás	31	3	1	Nem a Madridban focizott Beckham.
	INV szórend	37	19	2	Focizott Beckham a Madridban.
Többtagú tulajdonnevek	MFT	3	3	0	Megalakult az Arab Állami Egyetem.
	INV tagadás	0	1	1	Nem alakult meg az Arab Állami Egyetem.
	INV szórend	0	1	0	Az Arab Állami Egyetem megalakult.
Egymást követő azonos típusú tulajdonnevek	MFT	94	94	91	Gabi Gézától kapott egy csomagot.
	INV tagadás	1	3	6	Gabi Gézától nem kapott egy csomagot.
	INV szórend	1	0	4	Gabi Gézától egy csomagot kapott.
Szókincs	MFT	50	50	48	Szlovénia tengerparton helyezkedik el.
	INV tagadás	1	0	0	Szlovénia nem tengerparton helyezkedik el.
	INV szórend	2	7	0	Tengerparton helyezkedik el Szlovénia.

6 Diszkusszió

A legfontosabb következtetés, amit a 1. táblázatból levonhatunk, hogy míg mindhárom rendszer 95-97% F₁ értéket ér a SzegedNER korpusz tanító-kiértékelő részekre bontásán, a minimális működési tesztheink (MFT) felén nem megy át, még a legjobb névelem-felismerő rendszer sem (kilencből öt MFT teszt esetén hiba_{arány}(huBERT) $\geq \frac{1}{3}$). Azt is kijelenthetjük, hogy egyik rendszer sem képes kezelni az ‘egymást követő azonos típusú tulajdonnevek’ esetét.⁶ Hangsúlyozzuk, hogy a tesztek úgy állítottuk össze, hogy egyszerű, az ember számára egyértelmű feladatok legyenek, amelyeket

⁶ Megemlítjük ugyanakkor, hogy a szórendi variációk egyik esetét, amikor az ablativusban álló főnév előzi meg az alanyt (*Pétertől Évi kapott egy könyvet*), a huBERT már képes helyesen azonosítani, a másik két rendszernek azonban ez is nehézséget jelent.

minden névelem-felismerőnek illene teljesíteni (a szoftverfejlesztésben ez a *unit test*-nek felel meg). Ennek oka valószínűleg a nem megfelelő tanító adatbázis rendelkezésre állása, ugyanis a Szeged NE korpusz gazdasági rövidhírekből áll (Szarvas és mtsai, 2006b), míg a teszteseteink tartalmaznak hétköznapi életbeli (pl. *Megittam egy Sopronit*) és sport (pl. *Xavi a Barcelonában futballozott*) tematikájú mondatokat is.

Az invariancia teszteken (INV) azonban nagyon jól teljesít a huBERT, kijelenthetjük, hogy azokon mind átmegy (egyedül a toldalékolásos tesztek tagadásos variánsán változik az esetek több, mint 10 százalékában a predikció).

Ha a három rendszert összehasonlítjuk, akkor is a SzegedNER-es kiértékelésnél jóval árnyaltabb kép nyerhető az 1. táblázatból. A SzegedNER tanító-kiértékelő részekre bontásán alapuló kiértékelésekben a klasszikus gépi tanuláson alapuló rendszerek, mint a SzegedNER (Szarvas és mtsai, 2006a) vagy hunner (Varga és Simon, 2007) 95% körüli F_1 -értéket, míg a BERT alapú rendszerek - mind az mBERT, mind a huBERT - 97% körüli F_1 -értéket érnek el (Nemeskey 2020). Míg a SzegedNERen nincs szignifikáns különbség az mBERT és huBERT között, a fenti teszteken egyértelműen jobban teljesít a huBERT, hat MFT teszten felezi az mBERT hiba arányát és lényegében minden INV teszten átmegy, míg az mBERT-nél legalább négy esetben mondhatjuk, hogy elbukik (hiba_arány(mBERT) $\geq \frac{1}{3}$)).

Ha a tesztjeinken elért eredményeket vizsgáljuk, azt mondhatjuk, hogy az mBERT viselkedése közelebb áll a SzegedNERéhez, mint a huBERTéhez, ami ellentmond a SzegedNER korpuszon mért F_1 -értékek által festett képnek. Az mBERT csak a névutó MFT teszten teljesít jobban, mint a SzegedNER, igaz, robusztusabb a tagadás és szórendi változásokra (minden INV teszten, amin a SzegedNER elbukik, sokkal jobban teljesít az mBERT). Ez utóbbinak valószínűleg az a magyarázata, hogy a SzegedNER jellemzőkészletében fontos jellemzők az ún. ablakolt jellemzők, azaz pl. a címkézendő szót kettővel megelőző szó jellemzői, míg a BERT transzformer modellje az egész be-menetet figyelembe tudja venni.

Az eredmények részletesebb, nyelvi szinteket is figyelembe vevő elemzéséből az is kiviláglik, hogy - az egymást követő azonos típusú tulajdonnevek esetét leszámítva - a többes szám jelenségét, azaz egy morfológiai változást a legnehezebb kezelni a rendszereknek, hiszen itt láthatók a legmagasabb hibaarányok. Ezzel szemben egy másik morfológiai jelenség, a toldalékolás tesztjén viszonylag kevesebb hibát láthatunk: úgy tűnik tehát, hogy modelljeink fel vannak készítve a névelemek ragozott alakjainak kezelésére a magyar nyelvben. Érdekesképpen megjegyezzük, hogy míg utóbbi jelenség elsődlegesen a morfológiailag gazdag nyelvekre jellemző, addig a tulajdonnevek többes számba tétele (pl. márkanevek használata esetén) a nyelvek szélesebb körében ismert jelenség, így a jövőben mindenképpen hasznos lenne e nyelvi jelenségek vizsgálata más nyelvek CheckList-tesztjeiben is.

Ami a szintaktikai jellegű tesztekkel illeti, a névelő tesztjén rosszabbul teljesít a SzegedNER és az mBERT, mint a névutó esetében, a huBERT azonban azonos eredményt ér el. Úgy tűnik, az mBERT kevésbé ismeri fel a mondatkezdő pozícióban szereplő tulajdonneveket (pl. *Athén után ülésezett a NOB*), ami részben okozhatja a gyengébb teljesítményt a névutós tesztmondatok esetében.

A szemantikai jellegű tesztek esetében (predikátum cseréje, taxonómia) a SzegedNER és az mBERT egyaránt 50 körüli hibaarányt mutat, míg a huBERT 23-26-ot. Úgy tűnik tehát, hogy a szemantikai változásokra is robusztusabb a huBERT a másik két rendszernél. A szókincs tesztjén viszont mindhárom rendszernél gyakorlatilag azonos

hibaarányt láthatunk, noha jellemzően más hibákkal: míg az mBERT a terméknevek felismerésénél mutat hibákat, addig például a huBERT a szervezetnévként funkcionáló országneveknél hibázik többet.

Végül elmondhatjuk, hogy tesztheink közül kimagaslóan a legjobb teljesítményt érték el a rendszerek a többtagú tulajdonnevek azonosításában, minimális hibaarányokkal, ugyanakkor a legtöbb hibát pedig az egymást követő azonos típusú tulajdonnevek kezelésében érthetjük tetten. Ez arra utal, hogy a közvetlenül egymás mellett látott névelemek felismerése viszonylag nehéz feladat, egyben annak is jele, hogy mindegyik rendszer hajlamos az egymás mellett látott, azonos típusú névelemnek vélt elemet összevonni. Utóbbi sajátosság megint csak elsődlegesen a szabad szórendű (morfológiailag gazdag) nyelvekre jellemző, így hasonló nyelvek CheckList-es vizsgálata e téren is hozzájárulhat a névelem-felismerés kiértékelésének módszertani újragondolásához.

7 Összegzés

Cikkünkben bemutattunk magyar névelem-felismeréshez kilenc nyelvi jelenséget, amit 27 darab CheckList teszttel tudunk ellenőrizni. A három névelem-felismerő rendszer tesztelése fontos betekintést nyújt a rendszerek viselkedésébe.

Hangsúlyozzuk, hogy a CheckList kiértékelést kiegészítésként és nem alternatívaként, ajánljuk a klasszikus tanító- és kiértékelő adatbázisra bontáson számolt pontosság metrika mellett. Továbbá a kilenc nyelvi jelenség mellett, még számos más nyelvi jelenség tesztelhető a magyar névelem-felismerésben és minden feladat és alkalmazásnak saját nyelvi jelenségei vannak, azokat specifikusan kell definiálni. Cikkünk fő célja az, hogy minden olvasót motiváljunk arra, hogy értse meg jobban a nyelvtechnológiai alkalmazásainak viselkedését, amihez a CheckList keretrendszer egy hasznos eszköz.

Köszönetnyilvánítás

Farkas Richárd kutatási munkáját a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal Mesterséges Intelligencia Nemzeti Kiválósági Programja támogatta a 2018-1.2.1-NKP-2018-00008 azonosítójú projekt keretében.

Zahorszki Róbert munkáját a "Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein" című, EFOP-3.6.3-VEKOP-16-2017-0002 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

A publikációban szereplő kutatást az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

Hivatkozások

- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, Volume 58, pp 82-115 (2020)
- Devlin, J., Chang, M-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *NAACL*, pp. 4171–4186 (2019)
- Ethayarajh, K., Jurafsky, D.: Utility is in the Eye of the User: A Critique of NLP Leaderboards. In: *EMNLP* (2020)
- Hewitt, J., Manning, C.D.: A Structural Probe for Finding Syntax in Word Representations. In: *NAACL* (2019)
- Nemeskey D. M.: Egy `emBERT` próbáló feladat. In: *MSZNY* (2020a)
- Nemeskey, D. M.: *Natural Language Processing Methods for Language Modeling*. PhD disszertáció (2020b)
- Paroubek, P., Chaudiron, S., Hirschman, L.: Principles of Evaluation in Natural Language Processing. *Traitement Automatique des Langues, ATALA* 48 (1), pp.7-31 (2007)
- Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In: *ACL* (2020)
- Szarvas, Gy., Farkas, R., Kocsor, A.: A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: *Discovery Science, 9th International Conference*, pp. 268–278 (2006a)
- Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: Highly accurate Named Entity corpus for Hungarian. In: *International Conference on Language Resources and Evaluation* (2006b)
- Varga, D., Simon, E.: Hungarian Named Entity Recognition with a Maximum Entropy Approach. In: *Acta Cybernetica* 18(2), pp. 293–301 (2007)

Szerzői index, névmutató

- Ács, Judit, [15](#), [63](#)
- Bálint, Domonkos, [233](#)
- Balog, András, [139](#)
- Balogh, Réka, [147](#)
- Beke, András, [157](#)
- Berend, Gábor, [77](#), [91](#)
- Bial, Bence, [63](#)
- Bodor-Eranus, Eliza Hajnalka, [211](#)
- Csapó, Tamás Gábor, [123](#)
- Dömötör, Andrea, [243](#)
- Egas-López, José Vicente, [147](#)
- Farkas, Richárd, [197](#), [319](#)
- Fegyő, Tibor, [139](#)
- Feldmann, Ádám, [29](#)
- Ficsor, Tamás, [91](#)
- Főző, Eszter, [275](#)
- Galántai, Júlia, [211](#)
- Gosztolya, Gábor, [123](#), [147](#)
- Gulyás, Attila, [211](#)
- Gulás, Máté, [243](#)
- Hajdu, Róbert, [29](#)
- Halász, Dávid, [29](#), [39](#)
- Herdinai, Viktor, [197](#)
- Hoffmann, Ildikó, [147](#)
- Imre, Nóra, [147](#)
- Indig, Balázs, [29](#)
- Kalivoda, Ágnes, [39](#)
- Kálmán, János, [147](#)
- Kicsi, András, [275](#)
- Kilián, Imre, [265](#)
- Kornai, András, [15](#)
- Kulcsár, Virág, [53](#)
- Laki, László János, [243](#)
- Lévai, Dániel, [15](#), [53](#)
- Lipp, Veronika, [39](#)
- Makrai, Márton, [29](#)
- Markó, Alexandra, [123](#)
- Mihajlik, Péter, [139](#), [233](#)
- Mittelholcz, Iván, [29](#), [39](#)
- Nagy, Attila, [63](#)
- Nagy, Soma Bálint, [197](#)
- Nemeskey, Dávid Márk, [3](#), [15](#), [319](#)
- Novák, Attila, [107](#), [305](#)
- Novák, Borbála, [305](#)
- Nyéki, Bence, [291](#)
- Pákáski, Magdolna, [147](#)
- Ring, Orsolya, [211](#)
- Samu, Flóra, [211](#)
- Sass, Bálint, [29](#), [39](#)
- Shandiz, Amin, [123](#)
- Szabó, Martina Katalin, [169](#), [183](#), [211](#), [253](#)
- Szaszák, György, [157](#)
- Sztahó, Dávid, [157](#)
- Szvetelszky, Zsuzsanna, [211](#)
- Takács, Károly, [169](#), [211](#)
- Tarján, Balázs, [139](#)
- Tóth, László, [123](#), [147](#)
- Üveges, István, [169](#), [183](#), [211](#), [253](#)
- Váradi, Tamás, [29](#)
- Vidács, László, [275](#)

Vincze, Veronika, [147](#), [169](#), [183](#), [211](#),
[253](#), [275](#), [319](#)

Vit, Eszter, [211](#)

Yang, Zijian Győző, [29](#), [225](#), [243](#)

Zahorszki, Róbert, [319](#)

Zainkó, Csaba, [123](#)